ERASMUS UNIVERSITEIT ROTTERDAM

Thesis MSc.

Econometrics and Management Science

Quantitative Finance

Improving Value-at-Risk estimates by combining density forecasts

Author

Wanqian (Tony) Zhang[(a)]

Supervisor

Prof. dr. Dick van Dijk[(b)]

Co-reader

dr. Erik Kole[(b)]

(a) *Master of Quantitative Finance, Erasmus School of Economics*

(b) *Econometric Institute, Erasmus University Rotterdam*

March 2, 2013

Abstract

This research focuses on the properties of weighted linear combinations of prediction models, evaluated using log predictive scoring rule and new scoring rules based on conditional and censored likelihood for assessing the predictive accuracy of competing density forecasts over a specific region of interest, such as the left tail in financial risk management. We apply the technique above on 20 prediction models for forecasting the daily S&P 500 returns and analyze this framework both ex post and ex ante. We find that the VaR and ES estimates are more accurate through combining density forecasts using the conditional and censored likelihood scoring rules than the log predictive scoring rule.

Key words: Forecasting, Model combination, Density forecast evaluation, Scoring rules, Model Confidence Set, S&P 500 returns, Conditional likelihood, Censored likelihood, Risk management

Contents

1 Introduction

Managing uncertainty is of major importance for financial institutions and private clients worldwide. In financial risk management, Value-at-Risk (VaR) and Expected Shortfalls (ES) have become the standard measures of quantifying downside risk for investments. Hence, it is essential to develop approaches that provide accurate VaR and ES estimates. Over the last few years, *predictive densities* have received increasing attention in economics and finance. A density forecast is an estimate of the probability distribution of a random variable conditional on the information available at the time the forecast is made. Leading cases are in risk management as predictive density plays a central role for modelling VaR and ES. Furthermore, a density forecast has the advantage that it contains information on the uncertainties associated with for instance the conditional mean forecasts while a point forecast delivers no information about the characteristics of these uncertainties; e.g. see Granger and Pesaran [2000] and Garratt et al. [2003] as they argue that point forecasts are therefore rarely sufficient.

In the financial industry, many practitioners have developed their own risk management models for a wide range of applications. The growing evolution of regulation has further intensified the reliance on models but the use of these models can carry various types of model risk. As many risk measurement models have become increasingly complex, this has led to more exposure of model uncertainties which typically involves the possibility of incorrect business decisions or damage to the company's reputation. Above all, the credit crises in the recent years highlighted the significance of model risk. Thus, relying upon a single risk measurement model is dangerous.

In many situations, there may be multiple models available. Therefore, choosing one model and neglecting the other would lead to waste of information and it ignores the fact that the model may be among others misspecified, inaccurate in estimates of parameters or errors in assumptions. Combining different models we implicitly acknowledge that more than one model could provide good forecasts and we guard against misspecication by not putting all the weight on one single model. The intuition of optimal pooling is similar to that of portfolio optimization allowing the possibility that all of the models under consideration are false. Moreover, diversification gains can be achieved through optimal pooling by assigning positive weights to several models. It appears that a prediction model with less predictive accuracy can enter a optimal prediction pool if it outperforms other prediction models regularly.

A large econometric literature on forecast combinations of point forecasts can be found. Following from Bates and Granger [1969], forecast combinations has proven to be a highly successful forecasting strategy and has come to be viewed as a simple and effective way to improve the forecasting performance over the standard methodology. Examples of formal evaluations of forecast methods can be found in Stock and Watson [2004] include focus on macroeconomic forecasting, where forecast combination has performed well. Recent work of Rapach et al. [2010] shows combining forecasts deliver statistically and economically significant out-of-sample gains relative to the historical average consistently over time. Also see Timmermann [2006] for theoretical contributions on this field.

On the opposite, econometric literature on combinations of predictive density (optimal pooling) is much limited. The combination of density forecast is up to few years ago still an area waiting investigation. Wallis [2005] took the first steps into this unexplored area. He proposed a weighted linear combination of the competing density predictions resulting in a mixture distribution which provided us some key guidance in understanding combining density forecast. Yet, a little is known in how the weights on the competing density forecasts in the mixture should be determined. As Hall and Mitchell [2007] point out, the weights that we assign to the different competing density forecasts depend on how we decide to measure the accuracy of the resulting mixture predictive density. In their work, they propose a combining methodology which can be described as follows: Choose that set of weights to the competing density forecasts shaping the finite mixture that minimize the Kullback-Leibler information criterion. KLIC is the distance between the combined density forecast and the true but unknown density of the variable to be forecast. They prove that the optimal combined density obtained from this methodology cannot provide worst forecasts than the best individual forecast. Furthermore, KLIC is closely related to the logarithmic scoring rules, which are loss functions depending on the density forecasts and the true but unknown density aiming to evaluate accuracy. This nice relationship has contributed significantly to the increasing popularity of density forecasts and has led more works on evaluating accuracy in the scarce forecasting literature of combining density forecasts. In the recent work by Geweke and Amisano [2011], they have derived several interesting analytical results in the search of optimal pooling of density forecasts using log predictive scoring rule. They have shown that linear prediction pools could yield more accurate predictions as evaluated by a logarithmic score function. Among others, they have proved that how acknowledging that all the available models are false can result in improved predictions.

Techniques of evaluating density forecasts are developed at a high speed. One way is by comparing density forecasts relative to the data generating process which is discussed in Diebold et al. [1998]. In practice however, a caveat of this approach is that all models used to produce the density forecasts are "wrong". In fact, rejecting a model relative to the data generating process, which is called absolute evaluation, does not provide users enough information about the sufficiency of the model. For instance, if two models are both rejected (misspecified) or both accepted (correctly specified), then we have not enough evidence to prove which model should be preferred to the other. This issue has been put forward by Amisano and Giacomini [2007] and Giacomini and Komunjer [2005]. Another way of density forecast evaluation is by comparing competing density forecasts given a measure of accuracy. See e.g. Amisano and Giacomini [2007] and Bao et al. [2004] discussing the difference between the approaches and their preference of competing models. However, in applications when the set of competing density models are large, the search of the best single model is hardly realized because the data is not sufficient informative. Hansen et al. [2005]'s Model Confidence Set (MCS) provides convenient approach to this problem, which is in particular useful in applications without an obvious benchmark.

Financial institutions apply risk management to minimize and control the probability of certain unfortunate events causing losses. Appropriate preventing operations can be determined if practitioners have control over the outcome of these events. Hence, for practical reasons, we are especially interested in the probability of these events which can be translated into a particular region of the distribution - The LEFT TAIL. For instance, Amisano and Giacomini [2007] and Bao et al. [2004] suggest likelihood ratio tests based on KLIC-type logarithmic scoring rule aiming to evaluate and to compare density forecast over a relevant region. However, Corradi and Swanson [2006] and Gneiting and Ranjan [2008] correctly find that this approach is not appropriate for this task because the resulting predictive ability tests are biased toward densities with more probability mass in the region of interest. Diks et al. [2011] describe a possible solution to this problem by using new scoring rules based on conditional and censored likelihood for assessing the predictive accuracy of competing density forecasts over the region of interest. The underlying idea is to replace the full likelihood by the conditional likelihood, given that the actual observation lies in the region of interest, or by the censored likelihood, with censoring of the observations outside the region of interest.

In this paper we contribute to the literature on the comparative evaluation of combining density forecasts. We continue the work of Diks et al. [2011] and Geweke and Amisano [2011] on the class of Value-at-Risk estimation, which is in particular relevant for Managing Uncertainty. The underlying idea of our proposal is that we aim to improve the VaR estimates by combining density forecasts following the methods using KLIC-based scoring rules demonstrated by Geweke and Amisano [2011], extended with new scoring rules suggested by Diks et al. [2011] which make it possible to correctly evaluate accuracy in regions of interest comparing the relative performance of the competing combining density forecasts. From one thing to another, we further investigate on the methodology for achieving that set of weights to the competing density forecasts that maximize the newly proposed scoring rules. Alongside with this, the idea behind combining density forecasts will be motivated, used and strengthened in such a way that our resulting VaR and ES estimates are further improved in accuracy.

In our empirical application, we consider 4 types of volatility models and each has the choice of 5 different distributions of the innovations, allowing us to create 20 prediction models for forecasting the daily S&P 500 stock index return. Based on this research, several findings and future recommendations emerge: First, we show that both conditional likelihood and censored likelihood scoring rules are convenient metrics in comparing density forecasts when interest lies in a region instead of the whole distribution. Second, a higher score based on suitable scoring rules most likely leads to more appropriate VaR and ES estimates. Third, the choice of conditional distribution is more important than the choice of conditional volatility models in explaining the variability of density forecasts. Fourth, including relatively poor performers in pools of multiple models could lead to more accurate forecasts. We prove that our optimal pool outperforms the best individual models in terms of higher predictive scores. Fifth, by performing several back tests on the resulting VaR and ES estimates of the combined density forecasts, we demonstrate that the accuracy of these estimates are considerable improved. In addition, VaR and ES estimates are more accurate through combining density forecasts using conditional and censored likelihood scoring rules than logarithmic scoring rule.

This paper is organized as follows. We present the theoretical framework of combining density forecasts using scoring rules in section 2. Section 3 outlines an empirical application to investigate the adequacy of both individual prediction models and combined prediction models for the daily stock index return. Empirical results are reported and discussed in section 4. Section 5 concludes.

2 Combining density forecasts and scoring rules

In order to understand the intuition of combining density forecasts, we first explain the density forecast combination method proposed by Geweke and Amisano [2011] in section 2.1, based on the logarithmic scoring rule. As already mentioned, the logarithmic scoring rule is closely related to the KLIC and likelihood ratio tests, which are known to perform successful in many conventional statistical settings. In section 2.2 we provide alternative scoring rules suggested by Diks et al. [2011] for evaluating and comparing density forecasts in a specific region of interest. The main contribution of our paper is presented in 2.3, where we propose a methodology that aims to combine density forecasts by selecting the optimal weights based on the alternative scoring rules. In the remaining part of this section, we present both examples illustrating the intuition behind optimal pooling and pointing out the motivation of introducing the alternative scoring rules.

2.1 Combining densities using logarithmic scoring rule

In combining density forecasts, selecting the optimal set of weights on the competing density forecasts is essential. Broadly speaking, the way how we decide to measure the accuracy of the resulting mixture determines the construction of the optimal pool consisting of predictive densities.

Consider a vector time series $\mathbf{y_t}$, given its history $\mathbf{Y_{t-1}} = \{\mathbf{y_h}, ..., \mathbf{y_{t-1}}\}$, where h denotes the starting date of the time series and $h \leq 1$. A prediction model A constructs a predictive density for $\mathbf{y_t}$ with respect to an appropriate measure v from the history $\mathbf{Y_{t-1}}$. The predictive density takes the form $p(\mathbf{y_t}; \mathbf{Y^o_{t-1}}, A)$, where superscript "$o$" denotes the observed value. As Gneiting and Ranjan [2008] points out, the goal of density forecasting is to maximize the accuracy of the predictive distributions which is of major relevance in the financial industry.

In our study, we use past data alongside with scoring rules for assessing the performance of predictive densities. Scoring rules measure the quality of the density forecasts by assigning a numerical score based on the predictive distribution.

The logarithmic predictive score function of a single prediction model A is

$$LS(\mathbf{Y^o_T}, A) = \sum_{t=1}^{T} \log p(\mathbf{y^o_t}; \mathbf{Y^o_{t-1}}, A) \tag{2.1}$$

to assess the prediction performance of a model A over the sample period up to time T. This rule assigns a high score for the density forecast if the observation y_t falls within a region with high predictive density, and a low score if it falls within a region with low predictive density. In the literature, the logarithmic scoring rule is viewed as intuitively appealing and easy to interpret thanks to the following relationship with the goodness-of-fit measure Kullback-Leibler Information Criterion (KLIC), which measures the divergence of the density forecasts from the true density:

$$\begin{aligned} KLIC(A) &= E_t\left(\log p_t(\mathbf{y_t};\mathbf{Y^o_{t-1}}) - \log p_t(\mathbf{y_t};\mathbf{Y^o_{t-1}},A)\right) && (2.2)\\ &= \int_{-\infty}^{\infty} p_t(\mathbf{y_t};\mathbf{Y^o_{t-1}})\log\left(\frac{p_t(\mathbf{y_t};\mathbf{Y^o_{t-1}})}{p_t(\mathbf{y_t};\mathbf{Y^o_{t-1}},A)}\right)dy_t && (2.3)\end{aligned}$$

where $p_t(\mathbf{y_t};\mathbf{Y^o_{t-1}})$ denotes the true conditional density. Following equation (2.2), a higher logarithmic score is equivalent to a lower value of the KLIC. In practice however, the true conditional density is unknown. See e.g. Bao et al. [2004] where they verified a way to evaluate the KLIC indirectly. We can use the result of (2.2) to evaluate the relative accuracy of two competing densities by taking the difference between the $KLIC(A_i)$ and $KLIC(A_j)$, where $i \neq j$. This way, the true conditional density term drops out from (2.2). Moreover, the difference between the KLIC of the competing densities is equivalent to the difference of the logarithmic scores:

$$\begin{aligned} d_{ij}^{LS} &= LS(\mathbf{Y^o_T},A_i) - LS(\mathbf{Y^o_T},A_j) && (2.4)\\ &= \sum_{t=1}^{T}\log p(\mathbf{y^o_t};\mathbf{Y^o_{t-1}},A_i) - \sum_{t=1}^{T}\log p(\mathbf{y^o_t};\mathbf{Y^o_{t-1}},A_j), \qquad for\ i\neq j && (2.5)\end{aligned}$$

Amisano and Giacomini [2007] extend this methodology by proposing a weighted logarithmic scoring rule to focus on the performance of the density forecasts in the region of interest. The underlying idea is to emphasize and to compare the area of interest by applying a 'threshold' weight function. However, this scoring rule is biased in the sense that it gives higher scores to densities with more probability mass in the region of interest even if these densities are incorrect. We refer to the paper of Gneiting and Ranjan [2008] and Diks et al. [2011] where they illustrate such inconsistencies with some striking examples. A possible solution to this issue proposed by Diks et al. [2011] will be discussed in section 2.2. Turning to combining density forecasts, combination of probability densities $p(\mathbf{y_t};\mathbf{Y_{t-1}},A_i)$, where $i = 1,...,k$ takes the form

$$\sum_{i=1}^{k} w_i p(\mathbf{y_t};\mathbf{Y^o_{t-1}},A_i); \qquad \sum_{i=1}^{k} w_i = 1, \quad w_i \geq 0 \qquad (2.6)$$

Note that the restrictions are sufficient to ensure that (2.6) is a density function. As addressed by Geweke and Amisano [2011], this linear prediction pool is evaluated using the log predictive score function

$$\sum_{t=1}^{T} \log \left(\sum_{i=1}^{k} w_i p(\mathbf{y_t^o}; \mathbf{Y_{t-1}^o}, A_i) \right) \tag{2.7}$$

where T is sample size and the corresponding weight for each density forecast is determined based on the past performance of the pool. In this study, an optimal prediction pool is one with weights chosen so as to maximizes (2.7). As depicted by Hall and Mitchell [2007], the methodology for combining density forecasts aims to obtain the most accurate density forecast. They show that the maximization of (2.7) is a appealing way to evaluate density forecasts statistically. In addition, they compare the proposed methods with alternative evaluation methods in the literature, including probability integral transforms (PITS). Most simply involves the application of a Kolmogorov-Smirnov test for uniformity in many empirical studies. These alternative evaluation methods are mostly model-based and suffer from parameter uncertainty. For comparison discussions we refer to their paper. The focus of our work is in line with the methods of optimal pooling as proposed by Hall and Mitchell [2007] and Geweke and Amisano [2011].

2.2 Conditional likelihood and censored likelihood scoring rule

We consider new scoring rules based on conditional (*cl*) and censored likelihood (*csl*) proposed by Diks et al. [2011]. They have shown that these scoring rules are useful when the main interest lies in comparing the accuracy of density forecasts for a specific region, such as the left tail in financial risk management. For this purpose, the logarithmic scoring rule as outlined in the previous section does not satisfy this task. Diks et al. [2011] show that there can be incorrect density forecasts that receive a higher average score than the actual conditional density using this scoring rule. As a consequence, the outcome of the test of equal predictive accuracy could suggest incorrect density forecasts to be better than the true density. In their study they show through Monte Carlo simulations and an empirical application that *cl* and *csl* scoring rules have proven to be successful performers in forecasting of the true density. Intuitively, using these scoring rules to evaluate density forecasts' accuracy as a part of combining densities seems to be promising. In this section, we first describe these scoring rules.

We denote the likelihood-based scoring rules using the conditional likelihood as LS^{CL} for a specific region of interest B:

$$LS^{CL}(\mathbf{Y^o_T}, A) = \sum_{t=1}^{T} I(\mathbf{y^o_t} \,\epsilon\, B_t) \log \left(\frac{p(\mathbf{y^o_t}; \mathbf{Y^o_{t-1}}, A)}{\int_{B_t} p(s; \mathbf{Y^o_{t-1}}, A) ds} \right) \tag{2.8}$$

Where $I(\mathbf{y^o_t} \,\epsilon\, B_t)$ is an indicator function which takes the value 1 when the observed value falls within the region of interest B_t, or 0 otherwise. The *cl* rule allows us to evaluate the accuracy only on the specific region of interest by normalizing the density on the region of interest through $\int_{B_t} p(s; \mathbf{Y^o_{t-1}}, A)ds$. Furthermore, this normalization enables us to compare the competing density forecasts in terms of their relative KLIC values. Diks et al. [2011] argued that there is one caveat of the *cl* rule. Due to the normalization, it neglects the accuracy of the density forecast for the total probability of the region of interest. When the region of interest is the left tail, the *cl* cannot recognize different tail probabilities in density forecasts that have similar tail shapes. As a consequence, the *cl* scoring rule assigns comparable scores to predictive densities whether or not they match with the frequency at which tail observations actually occur. The (tail) probability is especially relevant in many risk management application and therefore it is of interest to includes this probability by introducing the censored likelihood (*csl*) scoring rule, denoted as LS^{CSL}

$$LS^{CSL}(\mathbf{Y^o_T}, A) = \sum_{t=1}^{T} I(\mathbf{y^o_t} \,\epsilon\, B_t) \log p(\mathbf{y^o_t}; \mathbf{Y^o_{t-1}}, A) + I(\mathbf{y^o_t} \,\epsilon\, B^c_t) \log \int_{B^c_t} p(s; \mathbf{Y^o_{t-1}}, A) ds \tag{2.9}$$

Where B^c_t is the complement of B_t. The censored likelihood scoring rule takes observations outside the region of interest into account but ignores the shape of the density outside B_t. In the similar way as before in (2.4), we can link the scoring rules and the KLIC such that the difference between the KLIC of the competing densities is equivalent to the difference of the scoring rules.

2.3 Combining densities using conditional likelihood and censored likelihood scoring rule

By extending the methodology discussed in the previous sections, we propose to combine density forecasts by selecting the optimal weights based on the new scoring rules.

First, combinations of probability densities $p(\mathbf{y_t}; \mathbf{Y_{t-1}}, A_i)$, where $i = 1, ..., k$, based on the conditional likelihood score LS^{CL} can be obtained in the similar way as in (2.6), given by

$$\sum_{i=1}^{k} w_i \, I(\mathbf{y_t^o} \,\epsilon\, B_t) \left(\frac{p(\mathbf{y_t^o}; \mathbf{Y_{t-1}^o}, A_i)}{\int_{B_t} p(s; \mathbf{Y_{t-1}^o}, A_i) ds} \right); \qquad \sum_{i=1}^{k} w_i = 1, \quad w_i \geq 0 \tag{2.10}$$

Once again, the restrictions used here are sufficient to ensure that (2.10) is a density function. $I(\mathbf{y_t^o} \,\epsilon\, B_t)$ is a indicator function which takes the value 1 when the observed value falls within the region of interest B_t, or 0 otherwise. Next, we introduce the evaluation function for combined densities in (2.10) based on the conditional likelihood log predictive score function, given by

$$\sum_{t=1}^{T} \log \left(\sum_{i=1}^{k} w_i \, I(\mathbf{y_t^o} \,\epsilon\, B_t) \left(\frac{p(\mathbf{y_t^o}; \mathbf{Y_{t-1}^o}, A_i)}{\int_{B_t} p(s; \mathbf{Y_{t-1}^o}, A_i) ds} \right) \right) \tag{2.11}$$

Which assigns a high score for the density forecast if the observation y_t^o falls within a region with high combined predictive density, and a low score if it falls within a region with low combined predictive density. In the same way, we can compute the combined probability densities based on the censored likelihood score LS^{CLS} by

$$\sum_{i=1}^{n} w_i \left(I(\mathbf{y_t^o} \,\epsilon\, B_t) p(\mathbf{y_t^o}; \mathbf{Y_{t-1}^o}, A_i) + I(\mathbf{y_t^o} \,\epsilon\, B_t^c) \int_{B_t^c} p(s; \mathbf{Y_{t-1}^o}, A_i) ds \right) \tag{2.12}$$

The corresponding censored likelihood log predictive score function is given by

$$\sum_{t=1}^{T} \left(\log \left(\sum_{i=1}^{k} w_i (I(\mathbf{y_t^o} \,\epsilon\, B_t) p(\mathbf{y_t^o}; \mathbf{Y_{t-1}^o}, A_i)) \right) + \log \left(\sum_{i=1}^{k} w_i (I(\mathbf{y_t^o} \,\epsilon\, B_t^c) \int_{B_t^c} p(s; \mathbf{Y_{t-1}^o}, A_i) ds) \right) \right) \tag{2.13}$$

Which enables us to evaluate the pools of densities based on the *cls* scoring rule.

For both *cl* (equation 2.11) and *csl* scoring rule (2.13), the weight vector is determined based on past data of the pool, updated at each t, recursively reflecting the accuracy of the prediction models in the pool predicting densities. Similar to log predictive scores, the optimal pool based on conditional and censored likelihood is the one that select that set of weights to the competing density forecasts maximizing equation 2.11 and2.13 respectively. This maximization is a convect programming problem which can be solved by using conventional software. In our study, we consider Matlab function fmincon.

2.4 Intuition behind combining density forecasts

In this section we present two examples to illustrate the idea of combining density forecasts. Recall from the previous sections, the computation of optimal weights which minimize the $KLIC$ distance between the combined and true densities plays a central role in optimal pooling. Furthermore, it is possible that by including inferior forecasts we can deliver more accurate density forecasts out of sample. In the first example, we show that there exist a set of weights in the combined pool that can beat the best individual predictions. Suppose we have two competitive prediction models A_1 and A_2 in the pool $[A_1, A_2]$. For $T = 3$, The values of the predictive densities $p(\mathbf{y_t}; \mathbf{Y^o_{t-1}}, A_i)$ are

Table 1: Predictive densities

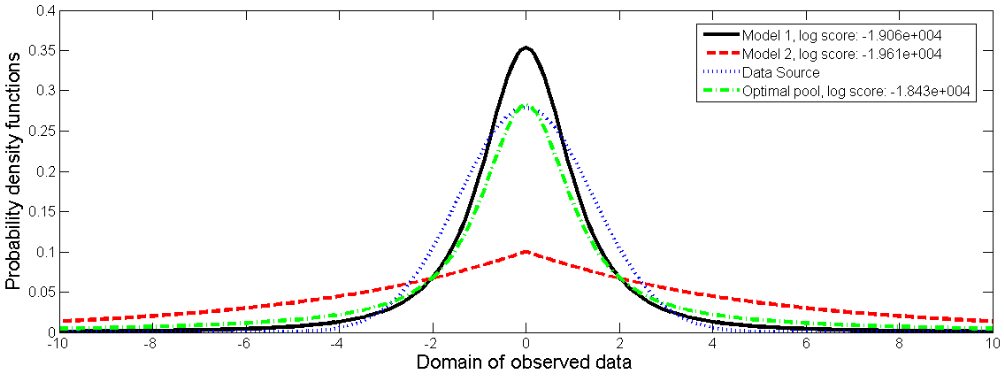
	A_1	A_2
$t = 1$	0.9105	0.3240
$t = 2$	0.7160	0.1228
$t = 3$	0.0348	0.9512

The log scores are $LS(\mathbf{Y^o_T}, A_1) = -3.7860$ and $LS(\mathbf{Y^o_T}, A_2) = -3.2742$. The optimal weights is $w^*_T = 0.6351$ such that model A_1 receives almost two-third of the weight despite of having a lower log score. In addition, the log score in the optimal pool takes the value of -2.0391 which beat the individual models.

Next, Figure 1 provides us a situation that further illustrates the intuition behind optimal pooling. In this example, two competitive prediction models are evaluated using the logarithmic scores with respect to the data generating process for $T = 200$. Clearly, we observe that model A_1 shows similar probability density function pattern (Student t distribution) as the data source (Normal distribution), while on the other hand model A_2, a flat probability density function (Laplace distribution) pattern, is not even close. The log score of model A_1 is therefore much higher. However, a surprising result emerges after we combine the models using optimal pooling. The combined model, denoted as ”Optimal pool”, closely tracks the data generating process. The highest log score is achieved for weights equal to 0.7210 and 0.2790 in this example. Moreover, even though model A_2 shows a lower log score, yet it receives a positive weight in the optimal pool. In sum, these examples suggest that by including relatively poor performers could lead to more accurate forecasts which strongly support the intuition behind optimal pooling.

Figure 1: Optimal pooling illustration

This figure presents an example of two competitive prediction models A_1 and A_2 evaluated using the logarithmic scoring rule for $T = 200$. The optimal weight is 0.7210 and 0.2790. The combined prediction model based on the competitive models is presented as "Optimal pool".



2.5 Motivation of introducing alternative scoring rules

In the next example, we demonstrate the problem that might occur by construction of the logarithmic scoring rule. As we will see, this scoring rule favors density forecasts with more probability mass in the region of interest over a less probability mass distribution, even if the latter is the true distribution from which the data is drawn. Suppose we once again have two competitive prediction models. This time, model A_1 is the Student-t distribution with v degrees of freedom, standardized to unit variance, with pdf

$$f(x|v) = \frac{\Gamma(\frac{v+1}{2})}{\Gamma(\frac{v}{2})} \frac{1}{\sqrt{\sigma^2(v-2)\pi}} \left(1 + \frac{)^2}{(v-2)}\right)^{-(\frac{v+1}{2})} \tag{2.14}$$

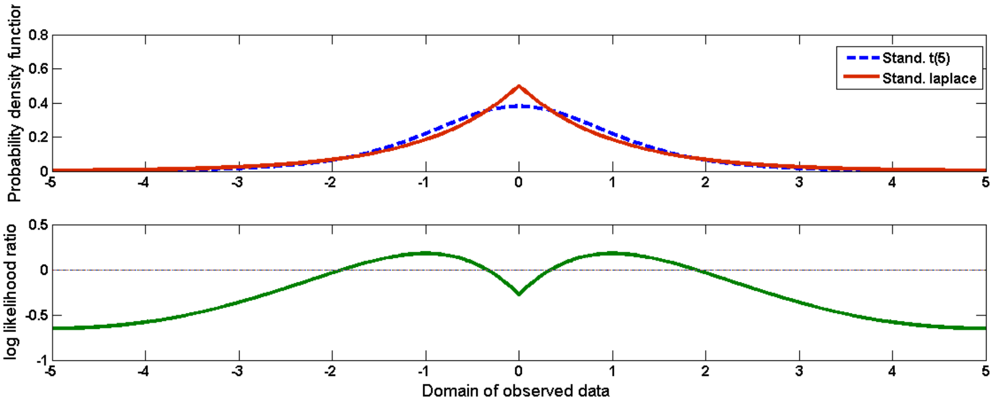
and model A_2 is the standardized Laplace distribution, with pdf

$$f(x) = \frac{1}{\sqrt{2}} \exp\left(-\sqrt{2}|x|\right) \tag{2.15}$$

The first plot of figure 2 shows the probability density functions for both models. In the second plot, the relative logarithmic likelihood score $LS(\mathbf{Y^o_T}, A_1) - LS(\mathbf{Y^o_T}, A_2)$ is displayed.

Figure 2: Logarithmic Scoring Rule

This figure presents an example of the probability density functions based on a Laplace and a Student t distribution in Panel 1. The relative accuracy score function between the competitive models is presented in Panel 2.

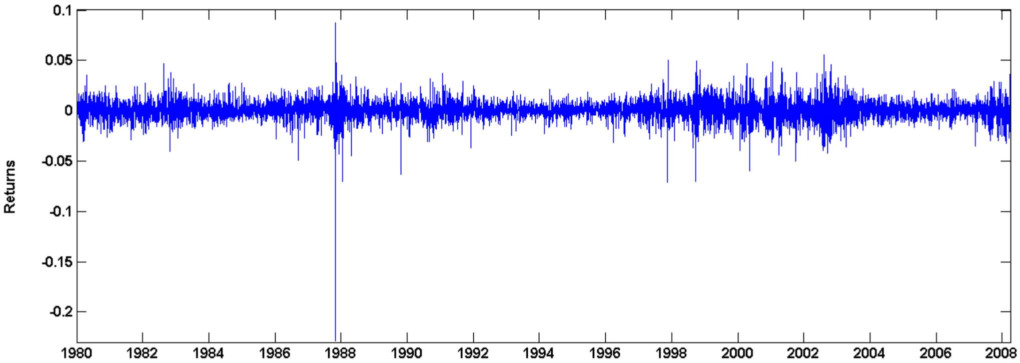


As the figure shows, the relative accuracy score function between the competitive models is negative for the region $(-\infty,-2)$ (Left tail) and $(2,+\infty)$ (Right tail) of the domain of observed data. Moreover, it appears that whenever there are observations in these regions the weighted logarithmic score difference is always negative. It reveals that the Laplace distribution is preferred over the Student-t distribution simply because the former has more probability mass in the regions of interest.

3 Application

In this section, we apply the proposed methodology to investigate the adequacy of both individual and combined density forecasts for the daily stock index returns. We consider S&P 500 log-returns $y_t = ln(p_t/p_{t-1})$, where p_t is the closing price on day t, adjusted for dividends and stock splits. The sample period runs from January 1, 1980 until March 14, 2008, giving us a total of $T = 7117$ observations.

Figure 3: S&P 500 log-returns



Let $\{y_t\}_{t=1}^{T}$ follow the stochastic process

$$y_t = \mu_t + \varepsilon_t = \mu_t + \sqrt{h_t}\eta_t \tag{3.1}$$

We consider an AR(5)[1] model for the conditional mean return μ_t, that is

$$\mu_t = \rho_0 + \sum_{j=1}^{5} \rho_j y_{t-j} \tag{3.2}$$

Next, a predictive density is based on two components: The specification of the distribution of the standardized innovations η_t and the specification of the volatility h_t.

For the specification of the innovations η_t we consider five candidate distributions. Four of them are symmetric parametric distributions: (i) Standard Gaussian normal (ii) Student t (iii) Generalized Error Distribution, GED (iv) Laplace and one skewed

[1] Regarding the order selection of "p" in the AR(p) model, we have considered Akaike information criteria (AIC) and Schwarz information criteria(BIC). Both model selection criteria show that the choice of p in our application has low sensitivity for the balance between goodness of fit and the number of parameters in the model. Therefore, we regard our assumption that the model order $p = 5$ as appropriate.

parametric distribution (v) Skewed Student t. The symmetric and skewed parametric distributions are described in more detail in the next section. They can be used in several settings, for instance the likelihood functions which we can apply in maximum likelihood estimation. Furthermore, the inverse CDF function can give the quantile of the standardized innovations, and the corresponding VaR of the return series.

The conditional variance h_t can be specified with various volatility models. A large literature on nonparametric, parametric and stochastic volatility models can be found. Given the possibility of hundreds of various GARCH-family models, We select four GARCH volatility models: (a) Symmetric normal GARCH(1,1) (b) Exponential GARCH(1,1) (c) Threshold GARCH(1,1) (d) Component GARCH(1,1). We discuss the specification of the volatility h_t in section 3.2. Furthermore, we consider daily models because the sign of the daily return has significant impact for future uncertainty. Returns tend to be more volatile after negative daily returns and less volatile after positive daily returns.

We apply a rolling window scheme for parameter estimation and evaluation of the prediction models. Let T be the total sample size and the length of the estimation window is set to $m = 2000$ observations. For each of the prediction model, on every time $t + m$ (t starting from 1 till $T - m + 1$) we calculate the maximum likelihood estimator (MLE) over historical data available at time $t+m-1$ and use that to obtain the predictive density. Our evaluation of the prediction models is based on their one-step-ahead density forecast of daily returns. The first one-step-ahead ahead forecasts are produced at time m, using data indexed $1, \dots, m$ and they are compared to y_{m+1}. The estimation window are then rolled forward one step and the second forecasts are obtained using observations $2, \dots, m+1$, and they are compared to y_{m+2}. This procedure is thus iterated, and the last forecasts are obtained using observations $T-m, \dots, T-1$, and they are compared to y_T. Normally, this yields a sequence of $n = T - m$ out-of-sample density forecasts. However, for real time forecasting, we consider a hold-out period for model construction. Moreover, we use the hold-out period of p observations for combining density forecasts such that the weights w of combined pools are determined. The remaining observations can then be used for evaluating the real time density forecasts. In this study, we set p equals to 1000, leaving us $T - m - p = 4117$ out-of-sample observations. Afterwards, predictive accuracy of the proposed volatility models are evaluated based on three scoring rules: (1) log predictive scoring rule, (2) conditional likelihood scoring rule and (3) censored likelihood scoring rule. For each of the scoring rules, we focus on the left tail of the distribution by using the threshold weight function $I(y \leqq \hat{r}_t^{\alpha})$. The threshold is time-varying and set equal to the empirical α-quantile of the return observations in the time varying estimation window. We consider $\alpha = 0.10, 0.05, 0.01$.

3.1 Specification of the distribution of the standardized innovations η_t

This section discusses the candidate distributions of the standardized innovations η_t included in this paper. The natural starting point in the literature is the standard Gaussian normal distribution due to its simplicity. However, from the stylized facts of the market return it has been shown that the distribution of the returns is non-normal, skewed and has excess kurtosis. Allowing innovations to be drawn by other distributions than normal can enhance the performance of GARCH models. A leptokurtosis distribution for example has a higher peak and greater mass in the tails than normal distribution of the same variance. Both leptokurtosis and negative skewness have impact on VaR. A commonly used alternative in the literature is the Student t-distribution which was first proposed by Bollerslev [1986], where the conditional distribution is assumed to be t-distributed. If we set the random variable X such that

$$X = \mu + \sigma T \tag{3.3}$$

Then X has a generalized Student t distribution with the distribution function given by

$$f(x|v) = \frac{\Gamma(\frac{v+1}{2})}{\Gamma(\frac{v}{2})} \frac{1}{\sqrt{\sigma^2(v-2)\pi}} \left(1 + \frac{(x-\mu)^2}{\sigma^2(v-2)}\right)^{-(\frac{v+1}{2})} \tag{3.4}$$

Where v is the degrees of freedom parameter, which is additionally estimated in the conditional variance equation alongside with other parameters. Furthermore, the degrees of freedom determines the excess kurtosis of the Student t distribution, which is equal to $6/(v-4)$, for $v > 4$. Therefore, the lower the degrees of freedom, the lower the peak of the distribution and the fatter the tail. Followed by the distribution function 3.4, the loglikelihood function of the Student t distribution can be derived which forms the basis for implementing maximum likelihood estimation

$$LLF = T \log\left[\frac{\Gamma(\frac{v+1}{2})}{\sqrt{\pi(v-2)}\Gamma(\frac{v}{2})}\right] - \frac{1}{2}\sum_{t=1}^{T} \log \sigma_t^2 - \frac{v+1}{2}\sum_{t=1}^{T} \log\left[1 + \frac{(x_t-\mu)^2}{\sigma_t^2(v-2)}\right] \tag{3.5}$$

Next, we introducte the Generalized Error Distribution (GED), which is a useful alternative and has been purposely applied in many applications such as modeling the stock market returns by financial corporations. The distribution function is given by

$$f(x|v) = v \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\lambda\sigma^2}\right)\left[\lambda\sigma^2 2^{\frac{v+1}{v}}\Gamma\left(\frac{1}{v}\right)\right]^{-1} \tag{3.6}$$

Where $\lambda = \left[\frac{\Gamma(\frac{1}{v})}{2^{2/v}\Gamma(3/v)}\right]$. The GED is flexible because it can be transformed through a parameter v into a distribution with fat tails ($v < 2$) or even into a platykurtotic distribution with thin tails ($v > 2$). When $v = 2$, the GED becomes a standard normal and when $v = \infty$, it becomes a uniform distribution.

A similar but different probability distribution to the GED is the Laplace distribution. Sometimes it is called the double exponential distribution. This distribution has more flexibility in the tails than a normal distribution in such a way that uncertainties of downside returns are incorporated more realistically. The generalized Laplace distribution function is given by

$$f(x) = \frac{1}{2\sigma}\exp(-\frac{|x-\mu|}{\sigma}) \tag{3.7}$$

All distributions mentioned above are categorized as symmetric distributions. Symmetric distribution have the common shortage that they cannot model skewness which might result in misinterpreting the risk. Therefore we need distributions that can capture skewness properly and these type of distributions are defined as skewed parametric distributions. The Skewed t distribution was first introduced by Hansen [1994] to model skewness in conditional distributions of financial returns by extending the Student t with a skew parameter. Since then, many other extensions of the Student t distributions have been proposed. For a detailed discussion of these distributions, see the review in Aas and Haff [2006]. In our study, we include for simplicity the Skewed t distribution in its basic form as our last candidate distribution. The pdf of the generalized Skewed t distribution consisting of one skewness parameter and two tail parameters is given by

$$f(x|\alpha, v_1, v_2) = \begin{cases} \frac{1}{\sigma}\left[1+\frac{1}{v_1}\left(\frac{x-\mu}{2\alpha\sigma K(v_1)}\right)^2\right]^{-(v_1+1)/2}, & \text{if } x \le \mu, \\ \frac{1}{\sigma}\left[1+\frac{1}{v_2}\left(\frac{x-\mu}{2(1-\alpha)\sigma K(v_2)}\right)^2\right]^{-(v_2+1)/2}, & \text{if } x > \mu. \end{cases} \tag{3.8}$$

where $\alpha \in (0,1)$ is the skewness parameter, $v_1 > 0, v_2 > 0$ are the left and right tail parameters respectively, $K(v) \equiv \frac{\Gamma((v+1)/2)}{\sqrt{\pi v}\Gamma(v/2)}$. We are aware of the fact that there are extensions of the Skewed t distribution which have been proven to perform well. For instance, as stated in As Zhu and Galbraith [2006], those could leads to a more adequately fit in the regions of interest such as the left tail. However, these computations are relatively time consuming due to the complexity of some of the extensions. A possible solution to this issue is by reducing the number of estimations through recursively determining the parameters for a larger interval between the estimation steps. For instance monthly instead of daily estimation. Despite the fact that choosing the most flexible distribution is not the focus of our research, we rightly think that more attention should be paid in the future on this area, and we acknowledge it as one of the limitations of this research.

3.2 Specification of the Volatility h_t

This section provides a description to the GARCH models that are used in this paper. The GARCH model is a generalization of the autoregressive conditional heteroscedasticity model (ARCH) which was first introduced by Engle [1982]. In the ARCH model, h_t is defined as the function of past squared errors. Bollerslev [1986] extended the ARCH model by including the lagged values of h_t and the GARCH model was born.

The natural benchmark model of the GARCH family is the symmetric normal GARCH(1,1) model, given by the following conditional variance equation

$$h_t = \omega + \alpha \varepsilon_{t-1}^2 + \beta h_{t-1} \tag{3.9}$$

Furthermore, we have to impose parameter constraints $\omega > 0$, $\alpha, \beta \geq 0$ to ensure that the conditional variance is finite and positive. Also the constraint $\alpha + \beta < 1$ is needed to achieve stationarity.

One of the strength of GARCH is that the parameters have a natural interpretation. (i) The GARCH error parameter α for instance measures the reaction of conditional volatility to market shocks. A large α might indicate a large sensitiveness of volatility to market events. (ii) The GARCH lag parameter β measures the persistence in conditional volatility. In case of a large β then volatility takes a relatively long time to vanish. (iii) The sum $\alpha + \beta$ is the rate of convergence of the conditional volatility which can be seen as the long term average level. (iv) $\omega/(1 - \alpha - \beta)$ determines the long term average volatility, which is the unconditional volatility in the GARCH model.

Among the large amount of variations in the selection of GARCH-family models, we have decided to involve the following candidates: GARCH(1,1) TGARCH(1,1), EGARCH(1,1), CGARCH(1,1). Each of them will be discussed next.

In practice, negative daily returns have larger impact than positive daily returns of the same magnitude. This feature is clearly observable in equity indices as future is more volatile after negative daily returns. Therefore, we allow this asymmetry by the GARCH model extended with a threshold term. The TGARCH(1,1) is given by the following equation

$$h_t = \omega + \alpha\varepsilon_{t-1}^2 I[\varepsilon_{t-1} \leq 0] + \gamma\varepsilon_{t-1}^2 I[\varepsilon_{t-1} \geq 0] + \beta h_{t-1} \tag{3.10}$$

where $I[A] = 1$ if A occurs, and 0 otherwise. In the same, parameter constrains are required for positiveness of h_t by $\omega > 0$, $\alpha > 0$, $\gamma > 0$, $\beta \geq 0$. And $(\alpha + \gamma)/2 + \beta < 1$ is needed for covariance stationarity.

The exponential GARCH was first introduced by Nelson(1991), this model formulate the conditional variance equation in terms of the *log* of the variance. It has the advantage that modelling *log* volatility no restrictions need to be imposed on parameters to ensure $h_t > 0$. The variance will always be positive even $log(h_t)$ is negative. The EGARCH(1,1) is given by

$$\ln(h_t) = \omega + \alpha_t z_{t-1} + \gamma_1(|z_{t-1}| - E[|z_{t-1}|]) + \beta_1 \ln(h_{t-1}) \tag{3.11}$$

where $z_{t-1} = \varepsilon_{t-1}/\sqrt{h_{t-1}}$. In the EGARCH model, he presence of leverage effects can be tested by the hypothesis that $\gamma < 0$. The impact is asymmetric if $\gamma \neq 0$

An alternative is the component GARCH model which was first introduced by Engle and Lee [1999]. The volatility in the CGARCH model is decomposed into a permanent or long-run and a transitory or short-run component. The transitory component is mean-reverting towards the trend component. One of the main reason to the improved performance is that the decomposition has led to more insight and flexibility in the explanation of persistency in the stock return volatility. For instance, the leverage effect has shown to be a short-run phenomenon and thus captured by the short-run component.

The conditional variance in the GARCH model shows mean-reversion to ω, which is a constant for all time. By contrast, the CGARCH(1,1) model allows mean-reversion to a varying level m_t, given by

$$h_t - m_t = \alpha(\varepsilon_{t-1}^2 - m_{t-1}) + \beta(h_{t-1} - m_{t-1}) \tag{3.12}$$

where m_{t-1} is the time-varying long-run volatility and given by

$$m_t = \omega + \rho(m_{t-1} - \omega) + \phi(\varepsilon_{t-1}^2 - h_{t-1}) \tag{3.13}$$

3.12 is the transitory component and converges to 0 with powers of $(\alpha + \beta)$. 3.13 is the long-run component and converges to ω with powers of ρ in the CGARCH model. By combining the transitory and the long-run equations we can rewrite the model by

$$h_t = (1-\alpha-\beta)(1-\rho)\omega+(\alpha+\phi)\varepsilon_{t-1}^2-(\alpha\rho+(\alpha+\beta)\phi)\varepsilon_{t-2}^2+(\beta-\phi)h_{t-1}-(\beta\rho-(\alpha+\beta)\phi)h_{t-2} \tag{3.14}$$

In summary, we can combine the candidate distributions and the volatility models as proposed in in section 3.1 and section 3.2 together giving us $5 * 4 = 20$ models in total. Therefore, we have created a collection of 20 density forecast models for improving the predictive accuracy through optimal pooling.

4 Empirical Results

In this section we report the empirical results applied on the daily S&P 500 return series. We start with discussing the predictive density accuracy for the individual density forecasts in section 4.1. We present predictive scores based on different metric of evaluation: (1) log predictive scoring rule, (2) conditional likelihood scoring rule (*cl*) and (3) censored likelihood scoring rule (*csl*) for all models and compare them first by taking the average score differences and we obtain test statistics based on the Diebold-Mariano tests of equal predictive accuracy (EPA). In order to compute this test, a benchmark model should be defined where the predictive score of this model is subtracted from the scores of alternative competing models. Unlike (EPA) type of tests, the Model Confidence Set (MCS) procedure as proposed by Hansen et al. [2005] does not require a benchmark to be specified, which is useful in our application of 20 prediction models where an obvious benchmark is difficult to be assigned. Also in line with the objective of this paper, additional attention is paid to the VaR and the ES estimates for different quantiles of the density forecasts. Afterwards, we continue with discussing the empirical results of the combined density forecasts and their predictive accuracy in section 4.2. We obtain the scores and the accompanying weights in the optimal pool first ex post. That is, we determine the optimal set of weights statically by looking back at the whole evaluation period. Despite the fact that this approach is not applicable in practice because only past data are available for optimization, this setting is however greatly illustrative and of importance in the sense that it offers us useful insights of how optimal pools can be constructed. Finally, realtime methods are applied and once again the scores and weights of the combined prediction models are reported in section 4.3. This time we determine the optimal set of weights dynamically given the information available at time t using a rolling window scheme. Note that for real time forecasting, we split the available data up to time t into two parts: An initial part that is used for model specification and parameter estimation; A second part, often termed as a hold-out period or a training period, for model construction. In this way, we dynamically obtain weights w of the opinion pools.

4.1 Individual prediction models

Table 2 presents the log predictive score for each model for the whole distribution evaluated over the out-of-sample period October 25, 1991 - March 14, 2008 (4117 observations). Results in this table are displayed as the sum of scores. The metric of evaluation assigns a high score for the density forecast if the observation y_t falls within a region with high predictive density, and a low score if it falls within a region with low predictive density.

Table 2: Predictive Density Evaluation - Log predictive scores for the whole distribution with number of out-of-sample observations n = 4117

This table presents the log predictive scores of the density forecasts which are obtained from 20 different prediction models: GARCH(1,1) and TGARCH(1,1) EGARCH(1,1) and CGARCH(1,1) with Gaussian Normal, Student t, GED, Laplace and Skewed t distribution, for daily S&P 500 returns over the evaluation period October 25, 1991 - March 14, 2008 (4117 observations). The scores presented in this table are the sum of scores over the evaluation period. Afterwards, the scores are sorted such that the ranking is reported between parentheses. That is, a ranking 1 corresponds with the highest score and ranking 20 the lowest.

GARCH - Normal	GARCH - Student t	GARCH - GED	GARCH - Laplace	GARCH - Skewed t
-5357 (18)	-5235 (3)	-5300 (13)	-5294 (12)	-5245 (6)
TGARCH - Normal	TGARCH - Student t	TGARCH - GED	TGARCH - Laplace	TGARCH - Skewed t
-5347 (17)	-5220 (1)	-5307 (15)	-5287 (11)	-5236 (4)
EGARCH - Normal	EGARCH - Student t	EGARCH - GED	EGARCH - Laplace	EGARCH - Skewed t
-5285 (10)	-5268 (8)	-5253 (7)	-5319 (16)	-5272 (9)
CGARCH - Normal	CGARCH - Student t	CGARCH - GED	CGARCH - Laplace	CGARCH - Skewed t
-5852 (20)	-5231 (2)	-5587 (19)	-5303 (14)	-5238 (5)

The highest predictive accuracy according to the log predictive score is obtained by the TGARCH model with Student t distributed innovations, exceeding the nearest competing model, CGARCH - Student t, by 11. The difference between the two models suggests a slight preference of the former model. Much clearer are the differences against the other models, yielding even greater favor of TGARCH - Student t. We also sort the scores from highest to lowest and the ranking are reported between parentheses. In this way, we may clarify the similarities or differences between the volatility models and the distribution specifications. Among the various candidate distributions, there is a strong evidence of superior predictive accuracy coming from the models that incorporate Student t, ranked 1, 2, 3 and 8, and Skewed t, ranked 4, 5, 6 and 9, distributed innovations. In addition, it reveals that the prediction models with Gaussian normal innovations perform the worst which is in line with our expectations. As the ranking displays, the performance of a prediction model depends more on the distribution specification, suggesting that the choice of conditional distribution is more important than the choice of conditional volatility models.

In Table 3 we present the predictive accuracy evaluated over the region of interest, for which $\alpha = 0.10, 0.05$ and 0.01, measured by different scoring rules. At first sight, the CGARCH family shows great dominance as the best performing predictive density method, yielding the highest scores for almost all quantiles considered, measured by all scoring rules. This outcome suggests that CGARCH may be seen as a flexible model in explaining the stock return volatility, through incorporating a transitory component which can adequately capture the short term features, such as the leverage effect. In addition, mainly CGARCH with Student t and Skewed t as candidate distributions outperform other competing models in terms of absolute evaluation. It is not surprising to observe that these candidate distributions generally perform well, not only for CGARCH, but for all volatility models considered in our collection of models. Clearly, the outcome of this analysis underlines the greater impact of the choice of conditional distribution above that of the choice of volatility models.

Furthermore, some interesting findings can be discovered within the Laplace distribution. Among the prediction models with Laplace distribution as candidate distribution, the first finding is the most obvious for CGARCH - Laplace. Reviewed by three scoring rules, CGARCH - Laplace shows overwhelming preference according to the log predictive scoring rule, ranked 2, 1, 1 respectively for $\alpha = 0.10, 0.05, 0.01$. In contrast, CGARCH - Laplace is only average based on the scores of both *cl* and *csl*, ranked between 9 and 11 for all region of interest which is clearly not favored over other competing models. For understanding this contradiction, it is useful to recap the shortcoming of log predictive scoring rule as we mentioned at the start of this study. When the main interest lies in comparing the predictive accuracy for a specific region, using logarithmic scoring rule is not appropriate for this task. This is because by construction, this metric tends to be biased in predictive ability toward densities with more probability mass in the region of interest, such as the Laplace distribution. Hence it is imaginable that the logarithmic scoring rule favors Laplace distribution only because of the ”fat” tails contained. This can be further clarified as Laplace distribution is apparently even more preferred by the logarithmic scoring rule for more extreme quantiles $\alpha = 0.05$ and 0.01, implying fatter tails in these regions. On the other hand, *cl* and *csl* scoring rules are by construction adjusted to this issue. Therefore, as Table 3 displays, Laplace distribution is clearly not preferred after normalizing in the region of interest (cl) and by taking density forecast outside the region of interest into account (csl),

Table 3: Predictive Density Evaluation - Predictive scores over the region of interest for $\alpha = 0.10, 0.05, 0.01$, with number of out-of-sample observations n $= 4117$

This table presents the scores based on (1) log predictive scoring rule (2) conditional likelihood scoring rule (3) censored likelihood scoring rule, over the region of interest for $\alpha = 0.10, 0.05, 0.01$, for daily S&P 500 returns over the evaluation period October 25, 1991 - March 14, 2008 (4117 observations). The scores presented in this table are the sum of scores over the evaluation period.

		Log predictive Scores			Conditional Likelihood Scores			Censored Likelihood Scores		
		$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
1	GARCH - Normal	-1191 (13)	-817 (18)	-330 (18)	-378 (18)	-235 (18)	-93 (18)	-1643 (17)	-1044 (18)	-381 (19)
2	GARCH - Student t	-1178 (10)	-771 (11)	-257 (11)	-304 (5)	-166 (2)	-44 (4)	-1561 (6)	-972 (6)	-312 (3)
3	GARCH - GED	-1207 (17)	-788 (14)	-279 (15)	-319 (14)	-183 (14)	-55 (14)	-1600 (14)	-1001 (14)	-335 (14)
4	GARCH - Laplace	-1177 (8)	-751 (4)	-241 (3)	-307 (10)	-170 (9)	-47 (10)	-1568 (8)	-977 (10)	-315 (8)
5	GARCH - Skewed t	-1164 (5)	-755 (6)	-250 (7)	-304 (6)	-167 (4)	-45 (5)	-1559 (5)	-971 (5)	-314 (7)
6	EGARCH - Normal	-1178 (11)	-803 (16)	-315 (17)	-361 (17)	-224 (17)	-84 (17)	-1611 (16)	-1015 (16)	-359 (17)
7	EGARCH - Student t	-1188 (12)	-776 (12)	-261 (12)	-305 (8)	-167 (5)	-44 (3)	-1569 (9)	-976 (9)	-315 (9)
8	EGARCH - GED	-1200 (15)	-781 (13)	-274 (13)	-312 (13)	-179 (13)	-53 (13)	-1579 (13)	-983 (13)	-324 (13)
9	EGARCH - Laplace	-1192 (14)	-760 (8)	-246 (5)	-308 (11)	-171 (11)	-48 (12)	-1574 (12)	-981 (11)	-317 (10)
10	EGARCH - Skewed t	-1176 (6)	-761 (9)	-255 (9)	-304 (7)	-168 (6)	-45 (6)	-1569 (10)	-975 (8)	-317 (11)
11	TGARCH - Normal	-1202 (16)	-825 (19)	-332 (19)	-381 (19)	-241 (19)	-96 (19)	-1644 (18)	-1043 (17)	-380 (18)
12	TGARCH - Student t	-1177 (9)	-769 (10)	-256 (10)	-303 (3)	-168 (7)	-45 (7)	-1556 (3)	-966 (1)	-310 (1)
13	TGARCH - GED	-1232 (19)	-805 (17)	-288 (16)	-323 (15)	-189 (16)	-58 (16)	-1607 (15)	-1004 (15)	-338 (15)
14	TGARCH - Laplace	-1176 (7)	-748 (3)	-239 (2)	-306 (9)	-171 (12)	-47 (11)	-1564 (7)	-973 (7)	-312 (4)
15	TGARCH - Skewed t	-1162 (3)	-752 (5)	-249 (6)	-303 (4)	-169 (8)	-45 (8)	-1554 (2)	-966 (2)	-312 (5)
16	CGARCH - Normal	-1315 (20)	-933 (20)	-390 (20)	-473 (20)	-293 (20)	-130 (20)	-1809 (20)	-1186 (20)	-444 (20)
17	CGARCH - Student t	-1162 (4)	-757 (7)	-250 (8)	-302 (1)	-165 (1)	-44 (2)	-1556 (4)	-968 (4)	-310 (2)
18	CGARCH - GED	-1224 (18)	-800 (15)	-278 (14)	-332 (16)	-183 (15)	-56 (15)	-1678 (19)	-1048 (19)	-340 (16)
19	CGARCH - Laplace	-1150 (2)	-731 (1)	-232 (1)	-308 (12)	-170 (10)	-46 (9)	-1571 (11)	-981 (12)	-317 (12)
20	CGARCH - Skewed t	-1145 (1)	-740 (2)	-244 (4)	-302 (2)	-166 (3)	-44 (1)	-1553 (1)	-967 (3)	-312 (6)

Note that the preceding scores here above are in terms of absolute evaluation, that is, on evaluating the given predictive accuracy, relative to the data-generating process. As already mentioned at the start of our study, for comparing competing models, a more practically interest is to discuss the relative score differences between the models in the decision making whether a model is preferred against another or not. Thus, it makes sense to consider average score differences for each scoring rule, denoted as S^*. First, the score difference is

$$d^*_{ij,t} = S^*[p(y^o_t; \mathbf{Y_{t-1}}^o, A_i)] - S^*[p(y^o_t; \mathbf{Y_{t-1}}^o, A_j)] \tag{4.1}$$

for $i \neq j$. Let $\overline{d}^*_{ij}$ denote the sample average of the score differences, that is,

$$\overline{d}^*_{ij} = n^{-1} \sum_{t=m}^{T} d^*_t \tag{4.2}$$

with $n = T - m$. The null hypothesis of equal scores is given by $H_{ij} : E(\overline{d}^*_{ij}) = 0$, for $t = m, m+1, ..., T-1$, and the alternative takes the form $E(\overline{d}^*_{ij}) \neq 0$. Naturally, Diebold-Mariano tests of equal predictive accuracy (EPA) can be applied. In order to compute this test, a benchmark model should be defined where the predictive score of this model is subtracted from the scores of alternative competing models.

However, it is essential to recall that a rejection of the EPA-test only identifies one or more models as significantly better than the benchmark. This type of tests provides little guidance about which models are the best performers that is relevant for optimal pooling. In addition, the choice of a certain model as benchmark is questionable in case of a large set of competing alternatives. Hansen et al. [2005]'s MCS approach has the advantage that a benchmark is not required. Another advantage of the MCS approach is that it acknowledges the limitations of the data. Unlike other model selection criteria, the MCS allows for the possibility that more than one model can be the best, in which case MCS contains more than a single model. Next, we briefly explain the intuition and general theory of MCS. For detailed discussion on this topic, we refer to Hansen et al. [2005].

The objective of the MCS procedure is to determine the set of models, denoted as M^*, that consists of the best model(s) from a collection of models given a level of confidence. In our application, the MCS is constructed from the collection of 20 competing models, denoted as M^0. These models are evaluated in terms of loss functions, in our case the different predictive density scoring rules. The MCS is based on an equivalence test, δ_M, where recursively models that are found to be significantly inferior to other models of M^0 are eliminated. An elimination rule, e_M identifies the model of M that has to be removed from M. We define the relative performance variables $d_{ij,t}$ and $\overline{d}_{ij}$ in a similar way as (4.1) and (4.2), where $\overline{d}_{ij}$ measures the relative sample loss between the i-th and j-th models. Next, the relative sample loss statistic to be used for the construction of MCS is defined as follows

$$\overline{d}_i \equiv m^{-1} \sum_{j \in M} \overline{d}_{ij} \tag{4.3}$$

which is the sample loss of the i-th model relative to the average across models in M. From this statistic we construct the t-statistic

$$t_i = \frac{\overline{d}_i}{\sqrt{\widehat{var}(\overline{d}_i)}} \tag{4.4}$$

where $\widehat{var}(\overline{d}_i)$ denote the estimate of $var(\overline{d}_i)$. This t-statistic is associated with the null hypothesis that $H_i = E(\overline{d}_i) = 0$ which forms the basis of tests of the hypothesis

$$H_{0,M} : E(\overline{d}^*_{ij}) = 0, \qquad \text{for all} \quad i, j \in M \tag{4.5}$$

The alternative hypothesis is denoted as $H_{A,M} : E(\overline{d}^*_{ij}) \neq 0$. Note that we take advantages of the equivalence between $H_{0,M}$, $\{H_{ij}$, for all $i, j \in M\}$, and $\{H_i$, for all $i \in M\}$. With $M = \{i_i, ..., i_m\}$ the equivalence follows from

$$E(\overline{d}_{i1}) = \cdots = E(\overline{d}_{im}) \Leftrightarrow E(\overline{d}_{ij}) = 0 \quad \text{for all } i, j \in M \Leftrightarrow E(\overline{d}_i) = 0 \quad \text{for all } i \in M \tag{4.6}$$

In order to test the hypothesis $H_{0,M}$, we apply the following test statistic,

$$T_{\max,M} = \max_{i \in M} t_i \tag{4.7}$$

With this test statistic and the associated $P_{H_{0,M}}$-value, the natural elimination rule is $e_{\max,M} \equiv \arg\max_{i \in M} t_i$. In this case, the elimination rule removes the model that contributes the most to the test statistic meaning that this model has the largest standardized excess loss relative to the average across all models in M. In addition, we introduce the MCS p-value as prescribed by Hansen et al. [2005]. This p-value is denoted as $\hat{p}_{e_{M_j}} \equiv \max_{i \leq j} P_{H_{0,M}}$. The MCS p-value has the advantage over the $P_{H_{0,M}}$ - value because it determines whether a model is in $M^*_{1-\alpha}$ or not, for any given α.

At last, the procedure of the MCS construction is as follows. We first set $M = M^0$. We test $H_{0,M}$ using δ_M at level α. If $H_{0,M}$ is accepted, we define the model confidence set $M^*_{1-\alpha} = M$, which consists the surviving models without being eliminated. Otherwise, we use e_M to eliminate a model from M and repeat the steps.

Next, we compute the MCS across all individual predictive density models. We define the MCS significance level α as 5%. Test results of the MCS are presented in Table 4. For each elimination step e_{M_j}, models that are eliminated and the associated MCS p-values are presented. Among the different scoring rules, only few models remain in the MCS based on the log predictive scoring rule. For $\alpha = 0.10$, after the elimination of 17 models, the MCS consists TGARCH - Skewed t, CGARCH - Laplace and CGARCH - Skewed t. Note that the MCS p-values cannot be interpreted as the probability that one of these models is the best model. For $\alpha = 0.05$ and 0.01, CGARCH - Laplace is the only surviving model. Turning to *cl* and *cls*, a first glance reveals that the resulting MCSs are consisted from a lot more of models, suggesting that the sample loss of a particular model relative to the average is smaller than using the log predictive scoring rule. Furthermore, we observe that TGARCH - Normal and CGARCH - Normal are consistently kicked out from the MCS based on the *cl* scoring rule. Another intriguing feature is that the smaller the region of interest, the larger the MCS in case of the *cls* scoring rule.

Table 4: Predictive Density Evaluation - Model Confidence Set for prediction models

In this table we present the MCS for individual prediction models given a MCS confidence level $\alpha = 5\%$. For each e_{M_j}, we report the to be eliminated model with the associated MCS p-value. Note that the following abbreviations are used for denoting the models: G = GARCH, E = EGARCH, T = TGARCH, C = CGARCH, . In each column, models above the line are eliminated while models under the line are the surviving models in the MCS

Elimination Rule	Log predictive Scores			Conditional Likelihood Scores			Censored Likelihood Scores		
e_{M_j}	alpha = 0.10	alpha = 0.05	alpha = 0.01	alpha = 0.10	alpha = 0.05	alpha = 0.01	alpha = 0.10	alpha = 0.05	alpha = 0.01
e_{M_1}	T-GED (0.003)	C-Norm (0.001)	G-Norm (0.003)	C-Norm (0.000)	T-Norm (0.000)	T-Norm (0.000)	C-Norm (0.002)	C-Norm (0.007)	C-Norm (0.045)
e_{M_2}	C-Norm (0.003)	T-Norm (0.001)	T-Norm (0.003)	T-Norm (0.000)	C-Norm (0.043)	C-Norm (0.033)	G-Norm (0.002)	C-GED (0.007)	T-Norm (0.045)
e_{M_3}	G-GED (0.003)	T-GED (0.001)	E-Norm (0.003)	G-Norm $\overline{(0.087)}$	G-Norm $\overline{(0.061)}$	E-Norm (0.033)	C-GED (0.004)	G-Norm (0.007)	G-Norm (0.045)
e_{M_4}	C-GED (0.003)	G-Norm (0.001)	C-Norm (0.003)	E-Norm (0.154)	E-Norm (0.089)	G-Norm (0.004)	T-Norm (0.004)	T-Norm (0.007)	E-Norm (0.045)
e_{M_5}	E-GED (0.003)	E-Stu*t* (0.001)	T-GED (0.006)	C-GED (0.280)	T-GED (0.164)	T-GED$\overline{(0.050)}$	G-GED (0.004)	E-Norm (0.009)	C-GED$\overline{(0.52)}$
e_{M_6}	E-Lap (0.003)	C-GED (0.001)	E-Stu*t* (0.006)	G-GED (0.485)	G-GED (0.170)	G-GED (0.050)	E-Norm (0.004)	G-GED (0.009)	G-GED (0.052)
e_{M_7}	E-Stu*t* (0.003)	G-Stu*t* (0.001)	G-GED (0.006)	T-GED (0.552)	C-GED (0.188)	E-GED (0.050)	T-GED (0.016)	T-GED (0.010)	T-GED (0.052)
e_{M_8}	T-Norm (0.004)	G-GED (0.001)	G-Stu*t* (0.006)	E-Lap (0.615)	E-GED (0.206)	C-GED (0.050)	E-Stu*t* (0.017)	E-Lap(0.010)	E-GED (0.052)
e_{M_9}	G-Stu*t* (0.004)	T-Stu*t* (0.001)	C-GED (0.006)	C-Lap (0.690)	E-Lap (0.210)	T-Lap (0.068)	E-GED (0.039)	E-GED $\overline{(0.058)}$	E-Lap (0.062)
$e_{M_{10}}$	T-Stu*t* (0.004)	E-Norm (0.001)	T-Stu*t* (0.006)	E-GED (0.764)	T-Lap (0.268)	E-Lap (0.068)	G-Lap (0.039)	G-Lap(0.107)	E-Ske*t* (0.071)
$e_{M_{11}}$	E-Ske*t* (0.004)	E-GED (0.001)	E-Ske*t* (0.008)	G-Lap (0.764)	G-Lap (0.315)	G-Lap (0.087)	E-Ske*t* (0.046)	C-Ske*t* (0.139)	G-Lap (0.130)
$e_{M_{12}}$	G-Lap (0.005)	E-Ske*t* (0.001)	E-GED (0.012)	T-Lap (0.853)	C-Lap (0.378)	C-Lap (0.135)	T-Stu*t* $\overline{(0.097)}$	E-Ske*t* (0.148)	C-Lap (0.171)
$e_{M_{13}}$	G-Norm (0.012)	E-Lap (0.001)	C-Stu*t* (0.012)	E-Stu*t* (0.875)	T-Ske*t* (0.378)	T-Ske*t* (0.141)	C-Lap (0.282)	E-Stu*t* (0.221)	E-Stu*t* (0.243)
$e_{M_{14}}$	T-Lap (0.012)	E-Stu*t* (0.001)	G-Ske*t* (0.014)	E-Ske*t* (0.891)	E-Ske*t* (0.380)	E-Ske*t* (0.167)	T-Lap (0.282)	T-Lap (0.250)	T-Lap (0.283)
$e_{M_{15}}$	G-Ske*t* (0.012)	-Ske*t* (0.002)	E-Lap (0.015)	G-Stu*t* (0.893)	E-Stu*t* (0.493)	G-Ske*t* (0.226)	G-Stu*t* (0.282)	G-Stu*t* (0.314)	T-Stu*t* (0.377)
$e_{M_{16}}$	E-Norm (0.025)	G-Lap (0.004)	T-Ske*t* (0.022)	G-Ske*t* (0.902)	G-Ske*t* (0.493)	T-Stu*t* (0.273)	G-Ske*t* (0.285)	G-Ske*t* (0.314)	G-Ske*t* (0.377)
$e_{M_{17}}$	C-Stu*t* (0.025)	T-Ske*t* (0.033)	C-Ske*t* (0.022)	T-Ske*t* (0.962)	T-Stu*t* (0.493)	E-Stu*t* (0.273)	E-Stu*t* (0.372)	T-Stu*t* (0.426)	C-Stu*t* (0.377)
$e_{M_{18}}$	T-Ske*t* $\overline{(0.091)}$	T-Lap (0.033)	G-Lap (0.022)	T-Stu*t* (0.962)	G-Stu*t* (0.504)	G-Stu*t* (0.273)	T-Ske*t* (0.712)	T-Ske*t* (0.794)	T-Ske*t* (0.377)
$e_{M_{19}}$	C-Lap (0.586)	C-Ske*t* (0.042)	T-Lap (0.030)	C-Ske*t* (0.962)	C-Ske*t* (0.504)	C-Ske*t* (0.287)	C-Stu*t* (0.712)	C-Stu*t* (0.794)	G-Stu*t* (0.377)
$e_{M_{20}}$	C-Ske*t* (1.000)	C-Lap $\overline{(1.000)}$	C-Lap $\overline{(1.000)}$	C-Stu*t* (1.000)	C-Stu*t* (1.000)	C-Stu*t* (1.000)	C-Ske*t* (1.000)	C-Ske*t* (1.000)	C-Ske*t* (1.000)

In order to determine whether a prediction model is accurate from a practical perspective, we apply out-of-sample forecast evaluation on the computation of VaR and ES estimates for each of the prediction models. When a target return is a α quantile of the return distribution, the probability of underperforming the target is α. If we know the CDF function of Y, then the corresponding quantile will be $x_t = F_y^{-1}(\alpha)$ such that the one-day VaR is determined by $P(Y_t < x_t) = \alpha$, where $VaR_{t,\alpha} = -x_t$. Furthermore, the expected shortfall, which is the expected loss given that the loss exceeds the VaR, is defined by $ES_{t,\alpha} = E(Y_t|Y_t \leq VaR_{t,\alpha})$.

Afterwards, model validation or backtesting approaches such as Christoffersen [1998]'s tests may be applied. These backtests use an i.i.d Bernoulli process, such that an exceedance of the VaR is tracked by an indicator function, typically $I(y_{t+1} \leq VaR_{t,\alpha})$ for $\alpha = 0.1$, 0.05 and 0.01, by assigning a value of 1 if the condition between parentheses is satisfied and 0 otherwise. We consider three types: (i) Unconditional coverage test (UC), (ii) Independence test (IND), (iii) Conditional coverage test (CC). Unconditional coverage tests are based on the number of exceedances, denoting times that the return y_t falls below the previous day's VaR estimate. We test the null hypothesis that the indicator function has a constant probability equal to the significance level of the VaR, α. The test statistic is a likelihood ratio statistic. Additionally, independence tests are used to test whether VaR exceptions are around the same time, commonly termed as clustering. In other words, when exceptions are not independent, an exceedance tomorrow is likely occur, given an exceedance today. We reject VaR models which exhibit clustered exceptions because this may indicate that the VaR model is not sufficient in changing market circumstances. The test statistic obtained from the independence tests is also a likelihood ratio statistic. For the sake of completeness, conditional coverage tests are formed by combining UC and IND into one test.

Another common used tail-related risk measure is the method developed by McNeil and Frey [2000] for backtesting $-VaR_{t,\alpha}$ and $ES_{t,\alpha}$ estimates, which is based on time series of standardized exceedance residuals, given by

$$\varepsilon_{t+1} = \begin{cases} \frac{y_{t+1} - ES_{t,\alpha}}{\hat{\sigma_t}}, & \text{if } y_{t+1} < -VaR_{t,\alpha}, \\ 0, & \text{otherwise.} \end{cases} \tag{4.8}$$

Here, $\hat{\sigma_t}$ is the one-day forecast of the standard deviation of the daily return obtained from the corresponding prediction model. The idea behind the test is based on the observation that, if the ES predictions are correct such that $ES_{t,\alpha}$ is an unbiased estimate for the expectation in the tail below the VaR, the expected value of ε_{t+1} should behave as a sample from an i.i.d. zero mean process. To elaborate on this, we test the null-hypothesis that ε_{t+1} has zero mean, against the alternative that the mean is positive(negative), such that the ES is too low(high) implying an overestimation(underestimation) of the ES. We consider for this purpose a two-sided t-test with a HAC variance estimator.

The results are summarized in Table 5. A glance at the backtesting results based on the coverages for each of the quantile display that the exceedance probabilities for most of the models are close to the empirical levels. For $\alpha = 0.10$, only volatility models with Gaussian normal innovations (model G-Norm, E-Norm, T-Norm and C-Norm) are rejected by the UC test at the 10% significance level, again indicating issues around these type of distribution in predictive accuracy. For $\alpha = 0.05$, this problem also seems to be the case for Laplace type of models (G-Lap, T-Lap, C-Lap), where the exceedance probabilities are clearly much smaller than the corresponding empirical levels. This points out that also Laplace type models might suffer from their specific distribution such that in some circumstances it would lead to misjudgement and possibly misspecfication on their VaR estimates. On the other hand, Student t and Skewed t type of models show good results according to the UC tests for all three quantiles, which again underlines their predictive strength. When we take a closer look to the outcome of IND tests, we also observe that even the most favored models from our collection of models sporadically fails to overcome clustering. In overall, Student t and Skewed t type models have achieved the most reliable VaR estimates according to CC tests. The outcome of these results illustrate the link between predictive scoring rules and accuracy of VaR estimates. Moreover, scoring rules can be seen as indicators such that models with higher scores imply potential in obtaining more accurate VaR estimates. Same conclusions can be drawn from the results of the ES estimates. To elaborate this, we firstly observe more extreme average ES estimates obtained by models with Laplace distribution. McNeil and Frey [2000] test rejects the null hypothesis at the 5% significance level for the regions $\alpha = 0.10$ and 0.01. In addition, the test statistics for this type of models are positive for $\alpha = 0.10$. Thus, this signifies that the ES estimates are too low and overestimated. On the other hand, we observe higher ES estimates for the normal distribution. For $\alpha = 0.10$, the null hypothesis is rejected for 3 out of 4 Normal models at the 10% significance level. Since the corresponding test statistics are negative, the ES estimates appear to be too high which also implies underestimation of the ES. Once again, this points out the inappropriateness of using normal distributed innovations for modeling stock returns. Finally, Student t, GED and Skewed t came out as more decent approaches according to McNeil and Frey [2000] tests, where null hypothesis is accepted for most of the volatility models with these distributions, for all three $\alpha's$. From these findings, we can confirm that models with high predictive power based on the proposed scorings rules also demonstrate positive outcomes in the ES predicitons. In other words, improving VaR and ES estimates can be achieved by developing methods that result in higher predictive density scores.

In summary, we find that both *cl* and *csl* scoring rules are convenient metrics in comparing density forecasts when interest lies in a region instead of the whole distribution. Furthermore, by applying different evaluation methods in comparing the prediction models, we have verified that a higher score in terms of predictive power according to *cl* and *csl* scoring rules indicates more accurate VaR and ES estimates. It is further of importance to point out that some of the models stand out compared to others, for instance Student t and Skewed t type models, but none of these models are consistently better. This suggests that relying too much on a single model is questionable. Thus, there is room for further investigation which leads us to the next stage of our research: Combining density forecasts.

Table 5: Predictive Density Evaluation - VaR and ES characteristics with number of out-of-sample observations n = 4117

This table summarizes the VaR and ES as risk measures with several additional backtesting approaches for each prediction models presented in columns. The rows are separated in three blocks, where each block corresponds with a region of interest given by quantiles. We consider three quantiles, 10, 5 and 1 ($\alpha = 0.10$, 0.05 and 0.01). The average VaRs reported are the observed average 1%, 5% and 10% quantiles of the density forecasts. The coverages correspond with the observed fraction of returns below the respective VaRs. The average ES values are equal to the conditional mean return, given a realization below the predicted VaR. Backtesting methods for VaR are labeled as UC, IND and CC. Here, we provide p-values for these tests. The last two rows for each block report McNeil-Frey test statistics and corresponding p-values for backtesting the ES estimates. Note that the numbers in the first row correspond with the prediction models similar to previous tables, where each model is assigned with a number.

		G-Norm	G-Stu*t*	G-GED	G-Lap	G-Ske*t*	E-Norm	E-Stu*t*	E-GED	E-Lap	E-Ske*t*	T-Norm	T-Stu*t*	T-GED	T-Lap	T-Ske*t*	C-Norm	C-Stu*t*	C-GED	C-Lap	C-Ske*t*
10	Av. VaR	-0.0120	-0.0110	-0.0110	-0.0112	-0.0111	-0.0117	-0.0109	-0.0108	-0.0111	-0.0110	-0.0119	-0.0110	-0.0107	-0.0113	-0.0111	-0.0127	-0.0112	-0.0121	-0.0118	-0.0113
	Coverage	0.0889	0.1049	0.1108	0.1008	0.1086	0.0896	0.1057	0.1091	0.1032	0.1076	0.0894	0.1023	0.1122	0.0979	0.1049	0.0821	0.1010	0.0923	0.0908	0.1018
	UC(p)	0.0837	0.4534	0.1046	0.9024	0.1946	0.0965	0.3899	0.1708	0.6224	0.2494	0.0984	0.7302	0.0648	0.7454	0.4534	0.0048	0.8731	0.2328	0.1549	0.7865
	IND(p)	0.4676	0.5133	0.5451	0.5370	0.5402	0.2440	0.0443	0.4340	0.0396	0.0565	0.2339	0.2459	0.9874	0.1046	0.1846	0.3054	0.1977	0.0806	0.1575	0.0593
	CC(p)	0.1721	0.6097	0.2231	0.8203	0.3574	0.1379	0.0915	0.2882	0.1066	0.0836	0.1257	0.4807	0.1818	0.2542	0.3131	0.0111	0.4306	0.1067	0.1339	0.1629
	Av. ES	-0.0165	-0.0171	-0.0178	-0.0181	-0.0169	-0.0166	-0.0170	-0.0173	-0.0181	-0.0170	-0.0163	-0.0171	-0.0172	-0.0179	-0.0169	-0.0164	-0.0179	-0.0174	-0.0178	-0.0167
	M-F	-1.7642	0.2413	2.3925	2.9832	-0.5913	-1.3465	-0.3125	0.9513	3.0123	-0.3129	-2.3215	-0.2911	0.6321	2.5632	-0.6231	-2.0345	2.4532	1.0652	2.4432	-1.1952
	M-F(p)	0.0778	0.8093	0.0168	0.0029	0.5543	0.1782	0.7546	0.3415	0.0026	0.7544	0.0203	0.7711	0.5274	0.0104	0.5332	0.0421	0.0142	0.2868	0.0146	0.2321
5	Av. VaR	-0.0155	-0.0149	-0.0151	-0.0162	-0.0152	-0.0152	-0.0148	-0.0148	-0.0159	-0.0151	-0.0154	-0.0149	-0.0146	-0.0162	-0.0151	-0.0165	-0.0152	-0.0167	-0.0170	-0.0156
	Coverage	0.0471	0.0539	0.0534	0.0413	0.0527	0.0491	0.0532	0.0530	0.0449	0.0551	0.0505	0.0561	0.0600	0.0410	0.0547	0.0476	0.0486	0.0481	0.0364	0.0466
	UC(p)	0.5402	0.4139	0.4734	0.0587	0.5713	0.8432	0.5049	0.5375	0.2777	0.2863	0.9125	0.2061	0.0408	0.0518	0.3337	0.6112	0.7635	0.6859	0.0027	0.4734
	IND(p)	0.2656	0.3039	0.3784	0.4334	0.4512	0.2607	0.0262	0.0889	0.1799	0.0288	0.2381	0.2366	0.0456	0.5798	0.1820	0.5332	0.2400	0.1518	0.0733	0.2446
	CC(p)	0.4461	0.4222	0.5247	0.1232	0.6414	0.5209	0.0677	0.1945	0.2258	0.0519	0.4957	0.2233	0.0167	0.1294	0.2571	0.7237	0.4792	0.3299	0.0022	0.3930
	Av. ES	-0.0199	-0.0205	-0.0205	-0.0218	-0.0208	-0.0199	-0.0205	-0.0204	-0.0215	-0.0207	-0.0192	-0.0198	-0.0213	-0.0214	-0.0206	-0.0197	-0.0212	-0.0211	-0.0226	-0.0206
	M-F	-1.2911	-1.1268	-1.1114	-0.6315	-1.0112	-1.2843	-1.1251	-1.1595	-0.7552	-1.0584	-1.4513	-1.3196	-0.8315	-0.7921	-1.0836	-1.3457	-0.8887	-0.9312	-0.2951	-1.0951
	M-F(p)	0.1967	0.2599	0.2664	0.5277	0.3119	0.1990	0.2606	0.2463	0.4502	0.2899	0.1468	0.1870	0.4057	0.4283	0.2786	0.1785	0.3739	0.3518	0.7679	0.2735
1	Av. VaR	-0.0221	-0.0241	-0.0239	-0.0275	-0.0248	-0.0216	-0.0239	-0.0233	-0.0271	-0.0246	-0.0220	-0.0238	-0.0230	-0.0274	-0.0244	-0.0235	-0.0248	-0.0264	-0.0287	-0.0256
	Coverage	0.0158	0.0104	0.0141	0.0053	0.0100	0.0160	0.0119	0.0138	0.0066	0.0114	0.0163	0.0104	0.0163	0.0063	0.0104	0.0197	0.0085	0.0148	0.0046	0.0080
	UC(p)	0.0137	0.8385	0.0754	0.0183	0.9848	0.0104	0.3938	0.0934	0.0899	0.5225	0.0079	0.8385	0.0079	0.0680	0.8385	0.0001	0.4775	0.0379	0.0054	0.3425
	IND(p)	0.5400	0.6124	1.0000	1.0000	0.5761	0.5569	0.7227	0.8702	0.3339	0.2704	0.5739	0.6124	0.2768	1.0000	0.2217	0.2455	1.0000	0.9432	1.0000	1.0000
	CC(p)	0.0397	0.8615	0.2057	0.0618	0.8551	0.0317	0.6528	0.2414	0.1489	0.4441	0.0251	0.8615	0.0163	0.1890	0.4642	0.0002	0.7770	0.1157	0.0209	0.6373
	Av. ES	-0.0259	-0.0285	-0.0261	-0.0349	-0.0281	-0.0255	-0.0278	-0.0275	-0.0335	-0.0261	-0.0260	-0.0276	-0.0258	-0.0345	-0.0274	-0.0231	-0.0304	-0.0252	-0.0373	-0.0297
	M-F	-0.2013	0.5984	-0.1351	2.3042	0.4711	-0.3329	0.3715	0.2846	1.9318	-0.1208	-0.1732	0.3154	-0.2411	2.1691	0.2492	-1.1132	1.1813	-0.4531	3.0859	0.9511
	M-F(p)	0.8405	0.5496	0.8925	0.0213	0.6396	0.7392	0.7103	0.7759	0.0534	0.9039	0.8625	0.7525	0.8095	0.0301	0.8032	0.2657	0.2375	0.6505	0.0021	0.3416

4.2 Combined prediction models - Ex post

Up to this point, individual prediction models have been investigated. We have seen that model with Student t and Skewed t distributed innovations have performed well. Our interest lies on the question whether these models are also present in pools of multiple prediction models and their ability to positively influence them.

The determination of weights depends on the choice of scoring rules. We consider log predictive, *cl* and *csl* scoring rules as described in equations 2.7, 2.11, and 2.13 for the pool optimization. Starting with the evaluation of the predictive densities for the entire out-of-sample period ex post. We assume in this case that we have access of the entire data up to the end of the outsample at a point of time of the sample period. Of course this scheme could not be used in practice since only past data are available. However, this setting is illustrative as a starting point in the large area of combining prediction models. In this way, we verify whether there exist a set of weights in the optimal pool which could outperform the best individual prediction models.

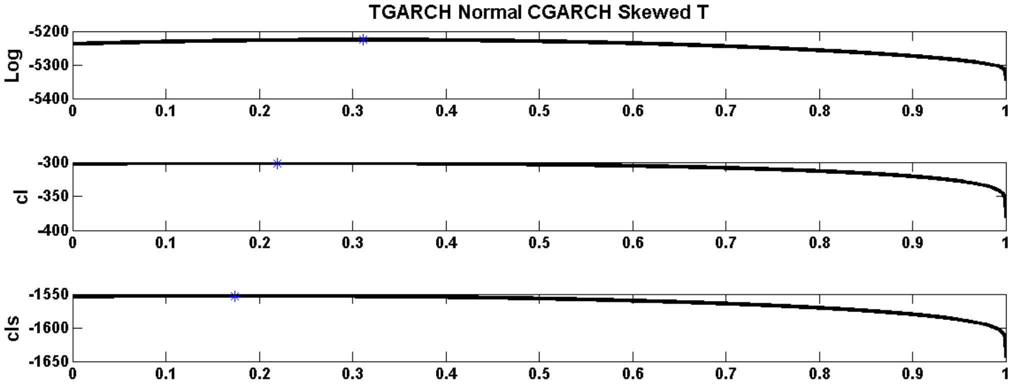
4.2.1 Pools of two models

Figure 4 shows a example of a pool of two models, TGARCH - Normal and CGARCH - Skewed t. The scores on the y-axis are presented as a function of the weights in the pool, presented on the x-axis. Recall from the previous section, the first model in this pool is known as not accurate whereas the second is proved as one of the best performing individual prediction model. Interesting result emerges when we combine these two models. First, a maximum log score of -5225 is achieved for $w = 0.31$, denoting the weight on TGARCH - Normal. Moreover, greater weight is given to the second model indicating that CGARCH - Skewed t is indeed favored in the optimal pool. The same holds true for *cl* where $w = 0.22$ with a score equals to -301, and for *csl* where $w = 0.18$ with a score equals to -1552. Second, even though we acknowledge that TGARCH - Normal has not shown sufficient predicting performance as a individual model, still the maximum log score obtained outperforms CGARCH - Skewed t's individual log score, as reported in 2, by 13. In addition, the highest individual *cl* and *csl* scores presented in 3 are both improved by 1 and 2 respectively.

We next generalize this methodology by combining the entire collection of 20 prediction models for pools of two models, giving us a total of 190 combined pools. Table 6

Figure 4: Predictive scores as a function of model weight

Scores are as a function ex post of the weights. Note that the x-axis is associated with the weight of the first mentioned model in the pool. For *cl* and *csl*, the scores correspond with the region of interest for α=0.10



reports the optimal log scores evaluated over the whole distribution and the corresponding weights for each of the combinations that can be created, table 7 and 8 present *cl* and *csl* scores for a pre defined region of interest $\alpha = 0.01$. Results for other regions, $\alpha = 0.10$ and 0.05 are reported in the Appendix. Each table consists a 20 by 20 matrix. Entries above the diagonal report the optimal scores achieved based on T observations. Entries below the diagonal correspond with the weight for models in the row of the table. Recall from the previous example, the optimal score for TGARCH - Normal with CGARCH - Skewed t pool is reported in cell row 11 and column 20, and the corresponding weight is reported in cell row 20 and column 11.

As demonstrated in Table 6, 33 out of 190 possible two model pools yield higher log predictive score in the optimum than the best individual model TGARCH - Student t, this may indeed give rise to possible improvement in forecasting by combining prediction models. Note that TGARCH - Student t itself is the most combined individual model for pools of two models. Moreover, this model shows great dominance in the optimal pools with other models as shown by the weights. However, the highest log predictive score of -5202 is achieved by combining the density forecasts of EGARCH - Normal and CGARCH - Skewed t, which outperforms TGARCH - Student t's log score by 18. Beforehand, it is imaginable that a Student t type model is included, but pointing EGARCH - Normal (only ranked 10 according to Table 2) as the other part in the best performing two model pool is not expected. This suggests that even individual models with poor performance could provide useful contribution in a pool of multiple model.

The same strategy can be applied on more convenient evaluation methods to verify combining models that could improve the scoring rules based on *cl* or *csl* on the region of interest, for instance the left tail. Table 7 reports combining *cl* scores for $\alpha = 0.01$. We observe 21 pools of two models yield higher score than the target score. These improvements are however small, due to the fact that for $\alpha = 0.10$ only 1/100 of the total observations are in this region. The number of evaluations by *cl* rule is therefore small, such that improvements are small. The best performer is the pool consisting CGARCH - Normal and CGARCH - Student t which outperforms the target score by 1.

From Table 8, we also observe that combining *csl* leads to more accurate density forecasts. Recall from the previous section, the target score *csl* score for $\alpha = 0.01$ equals to -310 obtained from TGARCH - Student t. Even though this model has proved to provide reliable predictions, still 41 combined models are able to beat this score. Once again EGARCH - Normal and CGARCH - Laplace is the most accurate out of 190 possible pools with a score equals to -304. Both models individually provide average or even poor predictions but together they form one of the most accurate combined model.

Table 6: Predictive Density Evaluation for combined models - Ex post optimal log predictive scores

In this table we presents the log predictive scores of the density forecasts for pools of two models. The scores presented in this table are the sum of scores over the evaluation period for the whole distribution evaluated over the out-of-sample period October 25, 1991 - March 14, 2008 (4117 observations). Entries above the diagonal are log scores of optimal pools. The corresponding weights of the pool for the model in that row are presented in entries below the diagonal. Note that the numbers in the first row and column correspond with the prediction models similar to previous tables, where each model is assigned with a number. The ”target score” is equal to -5220, referring to the best individual performance achieved by TGARCH - Student *t*. This target score is surpassed by 33 combinations of two models, as highlighted in grey. The best performing pool of two models consists EGARCH - Normal and CGARCH - Skewed *t*, which has achieved a score equals to -5202

	G-Norm	G-Stu*t*	G-GED	G-Lap	G-Ske*t*	E-Norm	E-Stu*t*	E-GED	E-Lap	E-Ske*t*	T-Norm	T-Stu*t*	T-GED	T-Lap	T-Ske*t*	C-Norm	C-Stu*t*	C-GED	C-Lap	C-Ske*t*
G-Norm		-5235	-5282	-5248	-5242	-5278	-5254	-5237	-5254	-5252	-5336	-5220	-5276	-5242	-5231	-5340	-5230	-5312	-5248	-5233
G-Stu*t*	0.95		-5235	-5234	-5233	-5207	-5235	-5213	-5235	-5234	-5229	-5219	-5231	-5231	-5223	-5234	-5228	-5234	-5232	-5227
G-GED	0.61	0.04		-5272	-5240	-5233	-5254	-5251	-5276	-5250	-5267	-5219	-5293	-5265	-5228	-5278	-5230	-5283	-5268	-5231
G-Lap	0.50	0.08	0.38		-5239	-5203	-5254	-5231	-5294	-5249	-5235	-5218	-5262	-5286	-5226	-5289	-5230	-5294	-5291	-5231
G-Ske*t*	0.82	0.29	0.76	0.77		-5207	-5241	-5212	-5241	-5245	-5233	-5218	-5232	-5234	-5234	-5243	-5227	-5243	-5235	-5236
E-Norm	0.88	0.51	0.65	0.67	0.54		-5221	-5225	-5213	-5218	-5284	-5207	-5239	-5206	-5210	-5268	-5203	-5250	-5206	**-5202**
E-Stu*t*	0.69	0.00	0.64	0.68	0.29	0.41		-5227	-5266	-5263	-5245	-5220	-5247	-5248	-5231	-5265	-5231	-5265	-5255	-5236
E-GED	0.72	0.50	0.90	0.81	0.55	0.51	0.64		-5237	-5222	-5238	-5210	-5253	-5233	-5211	-5232	-5211	-5237	-5230	-5208
E-Lap	0.45	0.04	0.29	0.00	0.17	0.31	0.14	0.13		-5264	-5243	-5219	-5268	-5287	-5230	-5308	-5231	-5317	-5301	-5236
E-Ske*t*	0.65	0.16	0.58	0.62	0.03	0.41	0.41	0.37	0.74		-5243	-5219	-5243	-5244	-5236	-5268	-5230	-5269	-5251	-5237
T-Norm	0.65	0.24	0.47	0.54	0.31	0.04	0.39	0.31	0.58	0.42		-5220	-5274	-5238	-5232	-5326	-5223	-5300	-5234	-5225
T-Stu*t*	0.96	0.84	0.92	0.90	0.79	0.60	1.00	0.63	0.94	0.89	0.99		-5220	-5218	-5219	-5217	-5215	-5217	-5214	-5213
T-GED	0.59	0.24	0.49	0.62	0.35	0.36	0.42	0.00	0.69	0.46	0.54	0.05		-5262	-5229	-5271	-5225	-5279	-5256	-5224
T-Lap	0.51	0.18	0.43	0.83	0.29	0.35	0.37	0.19	1.00	0.41	0.48	0.07	0.42		-5227	-5282	-5227	-5287	-5283	-5227
T-Ske*t*	0.80	0.49	0.72	0.75	0.68	0.51	0.71	0.51	0.81	0.91	0.78	0.21	0.72	0.75		-5233	-5218	-5234	-5222	-5228
C-Norm	0.02	0.01	0.10	0.12	0.02	0.03	0.03	0.08	0.17	0.05	0.04	0.01	0.13	0.12	0.02		-5231	-5492	-5299	-5237
C-Stu*t*	0.88	0.66	0.87	0.89	0.70	0.50	1.00	0.52	0.99	0.89	0.74	0.33	0.74	0.82	0.54	0.99		-5231	-5231	-5228
C-GED	0.13	0.01	0.06	0.00	0.01	0.10	0.03	0.05	0.07	0.05	0.14	0.01	0.10	0.00	0.02	0.62	0.01		-5303	-5237
C-Lap	0.46	0.14	0.35	0.40	0.25	0.31	0.30	0.18	0.77	0.36	0.43	0.14	0.36	0.32	0.26	0.88	0.06	0.97		-5234
C-Ske*t*	0.80	0.48	0.75	0.78	0.76	0.48	0.82	0.48	0.88	1.00	0.69	0.33	0.66	0.73	0.47	0.99	0.37	0.99	0.83	

Table 7: Predictive Density Evaluation for combined models - Ex post optimal conditional likelihood predictive scores for $\alpha = 0.01$

In this table we presents the conditional likelihood predictive scores of the density forecasts for pools of two models. The scores presented in this table are the sum of scores over the evaluation period for the regional distribution $\alpha = 0.01$, evaluated over the out-of-sample period October 25, 1991 - March 14, 2008 (4117 observations). Entries above the diagonal are log scores of optimal pools. The corresponding weights of the pool for the model in that row are presented in entries below the diagonal. Note that the numbers in the first row and column correspond with the prediction models similar to previous tables, where each model is assigned with a number. The "target score" is equal to -44, referring to the best individual performance achieved by CGARCH - Skewed t. This target score is surpassed by 21 combinations of two models, as highlighted in grey. The best performing pool of two models consists CGARCH - Normal and CGARCH - Student t, which has achieved a score equals to -43

	G-Norm	G-Stu*t*	G-GED	G-Lap	G-Ske*t*	E-Norm	E-Stu*t*	E-GED	E-Lap	E-Ske*t*	T-Norm	T-Stu*t*	T-GED	T-Lap	T-Ske*t*	C-Norm	C-Stu*t*	C-GED	C-Lap	C-Ske*t*
G-Norm		-44	-55	-47	-45	-83	-44	-53	-47	-45	-91	-45	-58	-47	-45	-88	-43	-56	-45	-44
G-Stu*t*	1.00		-44	-44	-44	-44	-44	-44	-44	-44	-44	-44	-44	-44	-44	-44	-43	-44	-44	-44
G-GED	1.00	0.00		-47	-45	-55	-44	-53	-47	-45	-55	-45	-55	-47	-45	-53	-43	-52	-46	-44
G-Lap	0.81	0.00	0.73		-45	-47	-44	-46	-47	-45	-47	-45	-47	-47	-45	-46	-43	-46	-46	-44
G-Ske*t*	1.00	0.00	1.00	1.00		-45	-44	-45	-45	-45	-45	-45	-45	-45	-45	-44	-43	-44	-45	-44
E-Norm	0.76	0.00	0.00	0.22	0.00		-44	-53	-47	-45	-84	-45	-58	-47	-45	-79	-43	-56	-45	-44
E-Stu*t*	1.00	0.00	1.00	1.00	1.00	1.00		-44	-44	-44	-44	-44	-44	-44	-44	-44	-43	-44	-44	-44
E-GED	1.00	0.00	0.78	0.37	0.00	1.00	0.00		-47	-45	-53	-45	-53	-47	-45	-52	-43	-51	-45	-44
E-Lap	0.82	0.00	0.73	0.00	0.00	0.81	0.00	0.62		-45	-47	-45	-47	-47	-45	-47	-43	-46	-46	-44
E-Ske*t*	1.00	0.00	1.00	1.00	0.00	1.00	0.00	1.00	1.00		-45	-45	-45	-45	-45	-45	-43	-45	-45	-44
T-Norm	0.38	0.00	0.00	0.14	0.00	0.00	0.00	0.00	0.12	0.00		-45	-58	-47	-45	-88	-43	-56	-46	-44
T-Stu*t*	1.00	0.00	1.00	1.00	0.45	1.00	0.20	1.00	1.00	0.89	1.00		-45	-45	-45	-45	-43	-45	-45	-44
T-GED	1.00	0.00	0.01	0.31	0.00	1.00	0.00	0.00	0.31	0.00	1.00	0.00		-47	-45	-54	-43	-53	-46	-44
T-Lap	0.77	0.00	0.67	0.00	0.00	0.76	0.00	0.62	0.52	0.00	0.83	0.00	0.67		-45	-46	-43	-46	-46	-44
T-Ske*t*	1.00	0.00	1.00	1.00	0.00	1.00	0.00	1.00	1.00	0.04	1.00	0.00	1.00	1.00		-45	-43	-45	-45	-44
C-Norm	0.34	0.06	0.16	0.26	0.11	0.30	0.07	0.15	0.25	0.10	0.37	0.09	0.16	0.29	0.14		-43	-56	-45	-44
C-Stu*t*	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.90		-43	-43	-43
C-GED	1.00	0.06	0.45	0.42	0.17	1.00	0.08	0.26	0.42	0.18	1.00	0.12	0.43	0.43	0.22	1.00	0.10		-45	-44
C-Lap	0.77	0.00	0.71	1.00	0.00	0.75	0.00	0.63	1.00	0.00	0.81	0.00	0.68	1.00	0.09	0.71	0.00	0.58		-44
C-Ske*t*	0.99	0.49	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.87	0.00	0.81	1.00	

Table 8: Predictive Density Evaluation for combined models - Ex post optimal censored likelihood predictive scores for $\alpha = 0.01$

In this table we presents the censored likelihood predictive scores of the density forecasts for pools of two models. The scores presented in this table are the sum of scores over the evaluation period for the regional distribution $\alpha = 0.01$, evaluated over the out-of-sample period October 25, 1991 - March 14, 2008 (4117 observations). Entries above the diagonal are log scores of optimal pools. The corresponding weights of the pool for the model in that row are presented in entries below the diagonal. Note that the numbers in the first row and column correspond with the prediction models similar to previous tables, where each model is assigned with a number. The ”target score” is equal to -310, referring to the best individual performance achieved by CGARCH - Student t. This target score is surpassed by 41 combinations of two models, as highlighted in grey. The best performing pool of two models consists EGARCH - Normal and CGARCH - Laplace, which has achieved a score equals to -304

	G-Norm	G-Stu*t*	G-GED	G-Lap	G-Ske*t*	E-Norm	E-Stu*t*	E-GED	E-Lap	E-Ske*t*	T-Norm	T-Stu*t*	T-GED	T-Lap	T-Ske*t*	C-Norm	C-Stu*t*	C-GED	C-Lap	C-Ske*t*
G-Norm		-312	-335	-313	-314	-356	-315	-324	-315	-316	-373	-310	-338	-310	-312	-359	-310	-331	-310	-312
G-Stu*t*	1.00		-312	-311	-312	-309	-312	-309	-311	-312	-312	-309	-312	-309	-310	-308	-310	-309	-309	-311
G-GED	1.00	0.00		-314	-314	-330	-315	-324	-315	-316	-334	-310	-333	-311	-312	-318	-310	-319	-312	-312
G-Lap	0.63	0.20	0.62		-312	-307	-312	-308	-315	-314	-311	-308	-311	-312	-309	-311	-310	-313	-314	-312
G-Ske*t*	1.00	0.00	1.00	0.73		-310	-313	-309	-313	-314	-313	-310	-312	-311	-311	-309	-310	-310	-311	-312
E-Norm	0.82	0.42	0.54	0.60	0.49		-311	-324	-309	-312	-359	-309	-333	-307	-311	-339	-307	-325	**-304**	-307
E-Stu*t*	0.89	0.00	0.95	0.59	0.31	0.54		-311	-314	-315	-314	-310	-314	-310	-311	-311	-310	-312	-311	-312
E-GED	1.00	0.52	0.87	0.72	0.60	0.90	0.62		-309	-311	-324	-309	-324	-308	-310	-310	-308	-312	-306	-308
E-Lap	0.60	0.13	0.51	0.00	0.17	0.40	0.23	0.25		-316	-312	-308	-312	-312	-310	-313	-310	-316	-316	-312
E-Ske*t*	0.81	0.00	0.75	0.45	0.00	0.48	0.00	0.33	0.68		-315	-310	-314	-311	-312	-312	-310	-313	-313	-312
T-Norm	0.59	0.15	0.30	0.47	0.25	0.00	0.28	0.04	0.48	0.34		-310	-337	-310	-312	-354	-309	-330	-308	-310
T-Stu*t*	1.00	0.78	1.00	0.84	0.89	0.82	0.93	0.78	0.87	1.00	1.00		-310	-307	-310	-304	-308	-305	-306	-309
T-GED	0.99	0.23	0.55	0.60	0.42	0.60	0.43	0.00	0.64	0.53	0.82	0.00		-310	-312	-317	-309	-318	-309	-310
T-Lap	0.68	0.37	0.71	1.00	0.51	0.46	0.54	0.38	1.00	0.73	0.61	0.20	0.53		-309	-308	-309	-311	-312	-311
T-Ske*t*	1.00	0.51	1.00	0.79	0.74	0.67	0.70	0.58	0.83	0.89	1.00	0.00	0.92	0.73		-306	-309	-307	-308	-310
C-Norm	0.23	0.15	0.17	0.36	0.20	0.19	0.19	0.14	0.37	0.23	0.25	0.13	0.14	0.33	0.18		-306	-333	-310	-308
C-Stu*t*	0.96	1.00	1.00	0.91	1.00	0.55	1.00	0.49	1.00	1.00	0.74	0.40	0.75	0.71	0.59	0.81		-308	-309	-310
C-GED	0.41	0.08	0.12	0.18	0.11	0.32	0.10	0.07	0.21	0.13	0.43	0.07	0.09	0.13	0.10	0.60	0.08		-315	-310
C-Lap	0.44	0.21	0.39	0.16	0.23	0.32	0.30	0.23	0.44	0.34	0.40	0.17	0.30	0.10	0.20	0.49	0.10	0.61		-311
C-Ske*t*	0.80	0.43	0.93	0.85	0.83	0.47	0.81	0.39	1.00	1.00	0.64	0.26	0.55	0.54	0.39	0.74	0.00	0.88	0.88	

4.2.2 Pools of multiple models

Next, we discuss the combining technique as described above to be applied on pools of multiple pools. It is well understood that the number of possibilities from combining 20 individual models is enormous. It would not make sense to report all of them. Instead, we consider segmentation of the individual models which will be demonstrated next. At the end of this section, we report the results for combining the whole collection of individual prediction models. Table 9 demonstrates optimal pooling using past data by combining all candidate distributions for each of the volatility model, giving us 4 new pools of models: 5-GARCH pool (including GARCH - Normal, GARCH - Student *t*, GARCH - GED, GARCH - Laplace and GARCH - Skewed *t*), 5-EGARCH pool, 5 TGARCH pool, 5-CGARCH pool. At the first sight, we observe small differences in scores between these pools, indicating that the resulting predictive accuracy between the volatility models are small given the same set of candidate distributions for combining. In addition, we can combine all volatility models for each of the candidate distributions. This time, allowing us to create 5 new pools of models: 4-Normal pool (including GARCH - Normal, EGARCH - Normal, TGARCH - Normal, CGARCH - Normal), 4-Student *t* pool, 4-GED pool, 4-Laplace pool, 4-Skewed *t* pool. We observe bigger range of the scores between the pools of models, which signifies the influence of candidate distributions.

As the results in table 9 display, the improvement in scores is present but small for *cl* and *csl* scoring rules compared to the individual and pools of two prediction models. This is due to the fact that our newly created pools are too restrictive such that neglecting other models would lead to waste of information. Therefore, it makes more sense to consider all models at the same time (20 in our case). In order to outperform the best individual models, we need to take a closer look into the construction of the pools of models.

Table 9: Predictive Density Evaluation - Ex Post - Pools of multiple models, with number of out-of-sample observations n = 4117

This table presents the scores of combined multiple models based on (1) log predictive scoring rule, over the whole distribution and (2) conditional likelihood scoring rule (3) censored likelihood scoring rule, over the region of interest for $\alpha = 0.10, 0.05, 0.01$, evaluated over October 25, 1991 - March 14, 2008 (4117 observations). The scores presented in this table are the sum of scores.

	Log predictive Scores	Conditional Likelihood Scores			Censored Likelihood Scores		
	Whole distribution	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
5 GARCH Models	-5238	-302	-166	-44	-1558	-970	-311
5 EGARCH Models	-5205	-298	-167	-44	-1552	-960	-308
5 TGARCH Models	-5217	-299	-167	-45	-1552	-963	-307
5 CGARCH Models	-5227	-301	-164	-43	-1552	-966	-306
4 Normal Models	-5264	-345	-208	-79	-1590	-996	-339
4 Student t Models	-5215	-301	-165	-43	-1552	-963	-308
4 GED Models	-5236	-302	-171	-51	-1565	-970	-311
4 Laplace Models	-5283	-306	-170	-46	-1563	-972	-312
4 Skewed t Models	-5228	-301	-166	-44	-1550	-963	-310

Table 10 confirms this phenomenon by reporting the average weights based on past data only for each of the individual models in the combined pools. Clearly, individual models that are found to be inferior according to our evaluation measures such as the MCS employ positive weights signifying substantial participation strength in the pools. In addition, Table 10 also provides insights about the amount of contributions from each individual models. Take EGARCH - Normal for example, this model shows the highest weight in the combined pool consisting only EGARCH models according to the *csl* scoring criteria for $\alpha = 0.10$, as well as the highest contribution in the combined pool consisting only Normal models. In other words, EGARCH - Normal can be seen as the biggest contributor, both in the EGARCH pool and the Normal pool of combined models. Based on this finding, we can conclude that this model might play an important role in the optimal pool. This strategy can be applied for all columns, such that it provides us some initial ideas about the combining power of each individual models before combining them all at once.

Table 10: Predictive Density Evaluation - Ex Post - Model weights -

This table presents the ex post optimal model. Rows in each block display the individual models forming that pool reported in Table 9 of combined models. Note that some model weights are highlighted, meaning that these models are both dominant in their volatility model family and distribution family.

		Log predictive Scores	Conditional Likelihood Scores			Censored Likelihood Scores		
		Whole distribution	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
5 GARCH Models	GARCH - Normal	0.13	0.00	0.00	0.00	0.00	0.12	0.00
	GARCH - Student t	0.00	0.30	0.90	1.00	0.00	0.23	0.80
	GARCH - GED	0.20	0.40	0.00	0.00	0.00	0.00	0.00
	GARCH - Laplace	0.02	0.30	0.10	0.00	0.15	0.19	0.20
	GARCH - Skewed t	0.66	0.00	0.00	0.00	0.85	0.46	0.00
5 EGARCH Models	EGARCH - Normal	0.45	0.06	0.00	0.00	0.47	0.51	0.37
	EGARCH - Student t	0.00	0.00	0.51	1.00	0.00	0.00	0.00
	EGARCH - GED	0.32	0.66	0.39	0.00	0.21	0.25	0.37
	EGARCH - Laplace	0.07	0.28	0.11	0.00	0.24	0.22	0.27
	EGARCH - Skewed t	0.16	0.00	0.00	0.00	0.08	0.03	0.00
5 TGARCH Models	TGARCH - Normal	0.03	0.00	0.00	0.00	0.00	0.02	0.00
	TGARCH - Student t	0.64	0.17	0.81	1.00	0.13	0.42	0.80
	TGARCH - GED	0.03	0.49	0.00	0.00	0.00	0.00	0.00
	TGARCH - Laplace	0.08	0.34	0.19	0.00	0.16	0.14	0.20
	TGARCH - Skewed t	0.22	0.00	0.00	0.00	0.71	0.42	0.00
5 CGARCH Models	CGARCH - Normal	0.01	0.00	0.23	0.10	0.05	0.05	0.19
	CGARCH - Student t	0.60	0.77	0.50	0.90	0.00	0.63	0.81
	CGARCH - GED	0.00	0.13	0.08	0.00	0.00	0.00	0.00
	CGARCH - Laplace	0.03	0.10	0.19	0.00	0.00	0.00	0.00
	CGARCH - Skewed t	0.36	0.00	0.00	0.00	0.95	0.32	0.00
4 Normal Models	GARCH - Normal	0.10	0.18	0.13	0.06	0.09	0.13	0.00
	EGARCH - Normal	0.88	0.71	0.61	0.67	0.83	0.81	0.81
	TGARCH - Normal	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	CGARCH - Normal	0.02	0.11	0.26	0.28	0.08	0.06	0.19
4 Student t Models	GARCH - Student t	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	EGARCH - Student t	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	TGARCH - Student t	0.67	0.49	0.00	0.00	0.55	0.64	0.60
	CGARCH - Student t	0.33	0.51	1.00	1.00	0.45	0.36	0.40
4 GED Models	GARCH - GED	0.06	0.16	0.23	0.25	0.07	0.08	0.14
	EGARCH - GED	0.89	0.75	0.58	0.47	0.87	0.88	0.80
	TGARCH - GED	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	CGARCH - GED	0.05	0.09	0.19	0.28	0.06	0.04	0.06
4 Laplace Models	GARCH - Laplace	0.00	0.00	0.32	0.00	0.00	0.00	0.00
	EGARCH - Laplace	0.00	0.11	0.00	0.00	0.00	0.00	0.00
	TGARCH - Laplace	0.68	0.68	0.00	0.00	0.76	0.84	0.90
	CGARCH - Laplace	0.32	0.21	0.68	1.00	0.24	0.16	0.10
4 Skewed t Models	GARCH - Skewed t	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	EGARCH - Skewed t	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	TGARCH - Skewed t	0.53	0.50	0.00	0.00	0.52	0.65	0.61
	CGARCH - Skewed t	0.47	0.50	1.00	1.00	0.48	0.35	0.39

At last, we discuss the results of combining the whole collection of individual prediction models. Table 11 reports scores and the weights in the optimal pool based on the scoring criteria for all three regions of interest. We observe that our optimal pool outperforms the best individual models as reported in table 2and table 3 For the log predictive score of the whole distribution, the score of the ex post optimal pool is -5194 against the best individual score -5220 obtained from Student t. The *cl* scores here are -295, -163, 43 against the best individual *cl* scores -302, -165, -44 for $\alpha = 0.10$, 0.05, 0.01 respectively. The *csl* scores here are -1544, -954, -303 against the best individual *csl* scores -1553, -966, -310. We recognize that the score differences decreases for smaller α's due to the number of observations within these areas. It is well understood that the smaller the number of to be combined predictive density forecasts, the smaller the improvement in the predictive density scores. Furthermore, The set of weights in the optimal pools presented here also confirms the strategy of determining the most influential models as proposed earlier. Moreover, the highlighted models in table 10 are indeed the biggest contributors in the optimal pool of 20 models.

In summary, the impact of combining is clearly present for log predictive score, but less observable for *cl* and *csl* scores. We have verified that there is a set of weights in the combined pool which can beat the best individual predictions. Furthermore, the results suggest that the inclusion of poor performers could lead to more accurate forecasts, which underlies the crucial intuition behind combining. From this, we can conclude that combining density forecasts not only provide us leads in terms of improvements to be achieved, but also surprising performances from individual ”out of consideration” models.

Table 11: Predictive Density Evaluation - Ex Post - Model Weights - Pool of full collection of models, with number of out-of-sample observations n = 4117

	Log predictive Scores	Conditional Likelihood Scores			Censored Likelihood Scores		
	Whole distribution	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
GARCH - Normal	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GARCH - Student t	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GARCH - GED	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GARCH - Laplace	0.03	0.00	0.00	0.00	0.00	0.00	0.00
GARCH - Skewed t	0.00	0.00	0.00	0.00	0.00	0.00	0.00
EGARCH - Normal	0.37	0.18	0.00	0.00	0.40	0.59	0.38
EGARCH - Student t	0.00	0.00	0.00	0.00	0.00	0.00	0.00
EGARCH - GED	0.25	0.53	0.25	0.00	0.01	0.02	0.14
EGARCH - Laplace	0.00	0.00	0.00	0.00	0.00	0.00	0.00
EGARCH - Skewed t	0.00	0.00	0.00	0.00	0.00	0.00	0.00
TGARCH - Normal	0.00	0.00	0.00	0.00	0.00	0.00	0.00
TGARCH - Student t	0.07	0.00	0.00	0.00	0.00	0.00	0.20
TGARCH - GED	0.00	0.00	0.00	0.00	0.00	0.00	0.00
TGARCH - Laplace	0.00	0.00	0.00	0.00	0.00	0.00	0.00
TGARCH - Skewed t	0.03	0.00	0.00	0.00	0.27	0.15	0.00
CGARCH - Normal	0.01	0.00	0.23	0.10	0.03	0.01	0.10
CGARCH - Student t	0.00	0.00	0.24	0.90	0.00	0.00	0.00
CGARCH - GED	0.00	0.00	0.02	0.00	0.00	0.00	0.00
CGARCH - Laplace	0.02	0.29	0.26	0.00	0.14	0.23	0.18
CGARCH - Skewed t	0.22	0.00	0.00	0.00	0.16	0.00	0.00
Score	-5194	-295	-163	-43	-1544	-954	-303

4.3 Combined prediction models - Real time

The optimal pooling procedure implemented in this section uses only the data available on each date, such that the weights in a optimal pool are continuously updated. For real time forecasting, we split the available data up to time t into an initial part that is used for model specification and parameter estimation, and a second part for model construction. Moreover, we make use of a hold-out period of p observations for combining density forecasts such that the weights w of combined pools are determined. The remaining observations are used for evaluaton. In our example, we define $p = 1000$. By letting the period length of 2000 observations for parameter estimation unchanged, this gives us 4117 observations for evaluation.

Real time forecasting results are summarized in Table 12, including predictive scores for pools of multiple models. Among others, we have evaluated pools of 4, 5 and 20 models similar to that in section 4.2. This time, optimal weights have been dynamically determined at each date for each combined pool. Next to this, we have also investigated the predictive accuracy of the equally weighted combined pool of 20 models. Note that this is similar to the "$1/N$" portfolio strategy in portfolio optimization which is commonly used as benchmark in forecasting. For comparison purposes, we have included one of the best individual model, CGARCH - Skewed t, based on our different evaluation criteria as reported in section 4.1. In addition, the inclusion of EGARCH - Normal and CGARCH - Laplace for comparison is obvious, as these candidate models together form the best performing combination among optimal pools of two models. According to the different scoring rules, our combined pool of 20 models is the most accurate. Diebold-Mariano test of equal predictive accuracy shows that the outperformance is only evidential in case of log predictive scores for whole distribution, whereas for smaller regions the differences are less clear. Furthermore, it is questionable whether combining all candidate models could lead to considerable improvement as in some cases pools of only two models show comparable predictive accuracy.

Once again, we can take a closer look into the construction of the optimal pools by analyzing individual model weights. Obviously, average weights over the evaluation period can be regarded as the most natural indicator to be used in order to measure the contribution of each individual model. In Appendix Table 19 and 20, the average model weights of combined pool consisting 4, 5 and 20 models are presented. First,

Table 12: Predictive Density Evaluation - Real Time - Pools of multiple models, with number of out-of-sample observations n = 4117

This table presents the scores based on (1) log predictive scoring rule, over the whole distribution and (2) conditional likelihood scoring rule (3) censored likelihood scoring rule, over the region of interest for $\alpha = 0.10, 0.05, 0.01$. We consider daily S&P 500 from January 1, 1980 until March 14, 2008, giving us a total of $T = 7117$ as sample period. For combined models, on each time t, weights of the pooling are determined dynamically through a rolling window scheme on past data using 1000 hold-out observations, after 2000 observations used for parameter estimation. The scores presented in this table are the sum of scores over the evaluation period with 4117 observations.

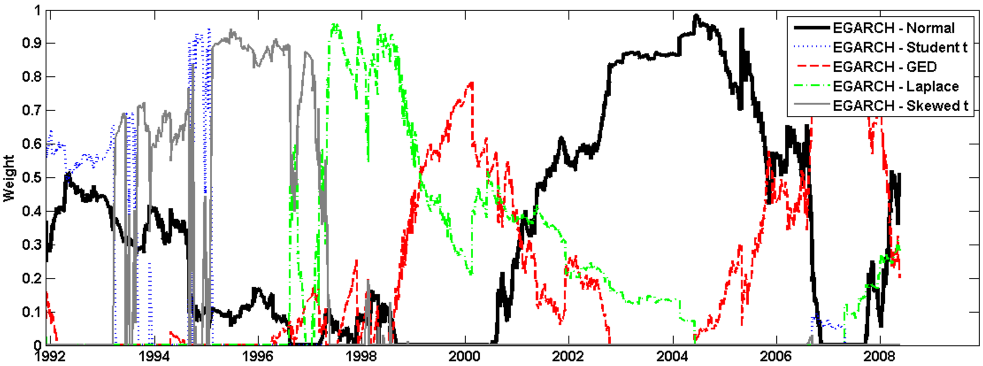
	Log predictive Scores	Conditional Likelihood Scores			Censored Likelihood Scores		
	Whole distribution	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
CGARCH - Skewed t	-5238	-302	-166	-44	-1553	-967	-312
EGARCH - Normal	-5285	-361	-224	-84	-1611	-1015	-359
CGARCH - Laplace t	-5303	-308	-170	-46	-1571	-981	-317
E-Norm * C-Lap	-5206	-297	-165	-45	-1546	-955	-304
5 GARCH Models	-5227	-312	-177	-54	-1569	-978	-328
5 EGARCH Models	-5197	-302	-168	-51	-1555	-968	-316
5 TGARCH Models	-5213	-311	-167	-56	-1557	-976	-327
5 CGARCH Models	-5231	-305	-165	-45	-1557	-971	-315
4 Normal Models	-5283	-359	-215	-85	-1608	-1017	-352
4 Student t Models	-5219	-299	-166	-42	-1554	-978	-311
4 GED Models	-5237	-313	-176	-55	-1571	-989	-323
4 Laplace Models	-5285	-307	-171	-47	-1566	-975	-313
4 Skewed t Models	-5233	-300	-166	-44	-1552	-966	-313
20 Equally Weighted	-5228	-300	-166	45	-1555	-964	-307
20 Dynamically Weighted	-5184	-298	-163	-42	-1546	-961	-301

we have noticed that the differences in weights between table 19, 20 and 10, 11 are small, suggesting that both dynamic and ex post weighting approach lead to similar optimal pools. Not surprisingly, our predictive scores as presented in table 12 by real time forecasting is close to that of table 9.

Furthermore, table 20 shows how the predictive performance of our total combined pool is realized. A notable observation is that a large part of our collection of individual models, around 80%, does not contribute in the optimal pool. These models are dominated by the other 20% of models, indicating that our collection of models are not fully utilized. This might explain the small improvement as diversification effect is not clearly present.

Figure 5: Optimal prediction pool weights for 5 EGARCH models, *csl* rule, $\alpha = 0.10$

This figure presents the daily updated weights for the combined pool which consists 5 EGARCH Models. Weights are determined by using the *csl* scoring rule for $\alpha = 0.10$. For each model, the average weight over the out-of-sample period (October 25, 1991 - March 14, 2008) are: EGARCH(1,1) - Normal, 0.34. EGARCH(1,1) - Student t, 0.08. EGARCH(1,1) - GED, 0.22. EGARCH(1,1) - Laplace, 0.20. EGARCH(1,1) - Skewed t, 0.16.



Instead of taking the average, we can directly investigate the evolution of model weights over time. As these change over time, we acknowledge that some models have more impact in certain periods. Figure 5 shows the optimal pool weights for combined 5 EGARCH models. From the start of our evaluation period up to around 1997, both EGARCH - Student t and EGARCH - Skewed t models dominate the optimal pool. This suggests that these models perform well in relatively less volatile periods of the financial market. Between 1997 and 2004, the world has faced several financial crises. Most prominent are the Asian financial crisis around 1998, the recession around 2000 and the bursting of the internet bubble. These events have led to a high volatile state of world. From 5, we observe that EGARCH - Normal, EGARCH - GED and EGARCH - Laplace have become more influential in the optimal pool during this relatively high volatile period. Among these models, the line representing the model weight for EGARCH - GED is the most notable. We can explain the evolution by analyzing the estimated shape parameter of the GED distribution. Between 1998 and 2002, the average value of this shape parameter v is 1.4, which theoretically represents a fat tail. After 2002, the shape parameter v is close to 2, which is closely to that of a normal distribution.

In the remaining part of this section, we evaluate the accuracy of our combined pools for each scoring rule in terms of VaR and ES estimates. Based on the daily updated weights of the optimal prediction pool, we can derive the predictive density function for the combined models subsequently compute the VaR and ES forecasts. Back testing on the resulting VaR and ES estimates are applied to compare the accuracy between the different scoring rules. Table 13 and 14 summarize the VaR and ES as risk measures with backtesting results for each combined prediction pools derived from logarithmic, *cl* and *csl* scores for 5 and 1 quantiles. Moreover, the derivation of the VaR and ES estimates for combined pools are based on the model weights as obtained from the optimal pooling methodology for each of the scoring rule. Therefore, different estimates result from different scoring for combined pools as different weights have been considered. Note that this is naturally not the case for equally weighted model and individual predictive models where we have only one unique VaR and ES estimate irrespective of the scoring rule. Furthermore, backtesting methods for VaR are labeled as UC, IND and CC recall from the previous sections with associated p-values for these tests. The last two rows report McNeil-Frey test statistics and the corresponding p-values for backtesting the ES estimates. The results show that empirical VaR exceedance probabilities are very close to the nominal levels for *cl* and *csl* scoring rules. Furthermore, both *cl* and *csl* are favored by the CUC test. Finally, McNeil-Frey test does not reject the *cl* and *csl* rules in approximately 90% of all cases. For all three quantiles, combined pool of 20 models shows the most reliable VaR and ES estimates resulting from combining through *csl* scores as suggested by the high p-values for UC and McNeil-Frey tests. In summary, the outcome of this analysis shows that the VaR and ES estimates are more accurate through combining density models using conditional and censored likelihood scoring rules than logarithmic scoring rule.

Table 13: Predictive Density Evaluation - Real Time - VaR and ES characteristics - Alpha = 0.05

This table summarizes the VaR and ES as risk measures with several additional backtesting approaches for each prediction model presented in columns. The rows are separated in three blocks, where each block corresponds with a region of interest given by quantiles. We consider three quantiles, 10, 5 and 1 (α = 0.10, 0.05 and 0.01). The average VaRs reported here are the observed average 5% quantiles of the density forecasts. The coverages correspond with the observed fraction of returns below the respective VaRs. The average ES values are equal to the conditional mean return, given a realization below the predicted VaR. Backtesting methods for VaR are labeled as UC, IND and CC. Here, we provide p-values for these tests. The last two rows for each block report McNeil-Frey test statistics and corresponding p-values for backtesting the ES estimates.

		C-Sket	E-Norm	C-Lap	E-Norm * C-Lap	5-GARCH	5-EGARCH	5-TGARCH	5-CGARCH	4-Normal	4-Student t	4-GED	4-Laplace	4-Skewed t	Equal	20Full
CON	Av. VaR	-0.0156	-0.0152	-0.017	-0.0154	-0.0155	-0.0153	-0.0153	-0.0168	-0.0155	-0.0149	-0.015	-0.0164	-0.0153	-0.0156	-0.0162
	Coverage	0.0466	0.0491	0.0364	0.0476	0.0471	0.0486	0.0496	0.0374	0.0471	0.0537	0.0547	0.0406	0.0513	0.0447	0.0391
	UC(p)	0.473	0.843	0.003	0.611	0.540	0.763	0.924	0.006	0.540	0.443	0.334	0.031	0.793	0.255	0.055
	IND(p)	0.245	0.261	0.073	0.202	0.125	0.113	0.138	0.092	0.266	0.030	0.128	0.228	0.269	0.113	0.411
	CC(p)	0.393	0.5209	0.0022	0.389	0.255	0.272	0.332	0.005	0.446	0.071	0.197	0.0486	0.525	0.149	0.113
	Av. ES	-0.0206	-0.0199	-0.0226	-0.0208	-0.0215	-0.0211	-0.0209	-0.0232	-0.0212	-0.0224	-0.0204	-0.0219	-0.02	-0.0221	-0.0224
	M-F	-1.0951	-1.2843	-0.2961	-1.0311	-0.7468	-0.8976	-0.9948	-0.1358	-0.8582	-0.4008	-1.1732	-0.6097	-1.2877	-0.5286	-0.4018
	M-F(p)	0.2735	0.199	0.7679	0.3025	0.4552	0.3694	0.3199	0.8920	0.3908	0.6886	0.2408	0.5421	0.1979	0.5971	0.6878
CEN	Av. VaR	-0.0156	-0.0152	-0.017	-0.0154	-0.0153	-0.0151	-0.0151	-0.0168	-0.0154	-0.015	-0.015	-0.0164	-0.0153	-0.0156	-0.0156
	Coverage	0.0466	0.0491	0.0364	0.0471	0.0496	0.051	0.0522	0.0386	0.0474	0.0539	0.0542	0.0406	0.0522	0.0447	0.0454
	UC(p)	0.473	0.843	0.003	0.540	0.924	0.832	0.642	0.013	0.575	0.414	0.386	0.04	0.642	0.255	0.43
	IND(p)	0.245	0.261	0.073	0.266	0.138	0.085	0.075	0.191	0.384	0.032	0.232	0.269	0.229	0.113	0.383
	CC(p)	0.393	0.521	0.002	0.446	0.332	0.223	0.184	0.019	0.585	0.072	0.336	0.066	0.435	0.149	0.500
	Av. ES	-0.0206	-0.0199	-0.0226	-0.0208	-0.0208	-0.0206	-0.0203	-0.0229	-0.021	-0.0224	-0.0205	-0.0216	-0.0197	-0.0221	-0.021
	M-F	-1.0951	-1.2843	-0.2961	-1.0285	-1.0235	-1.0954	-1.2033	-0.2018	-0.9384	-0.4056	-1.1388	-0.7158	-1.3589	-0.5286	-0.9418
	M-F(p)	0.2735	0.199	0.7679	0.3038	0.3061	0.2734	0.2289	0.8401	0.3481	0.6851	0.2548	0.4935	0.1742	0.5971	0.3463

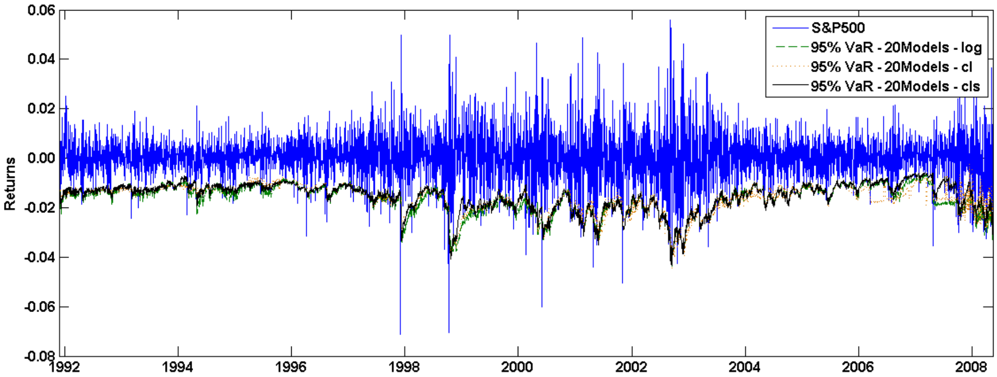
Table 14: Predictive Density Evaluation - Real Time - VaR and ES characteristics - Alpha = 0.01

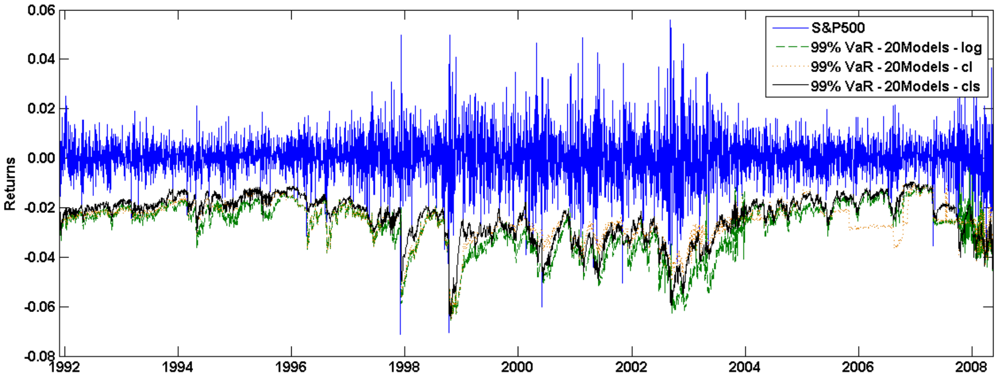
This table summarizes the VaR and ES as risk measures with several additional backtesting approaches for each prediction model presented in columns. The rows are separated in three blocks, where each block corresponds with a region of interest given by quantiles. We consider three quantiles, 10, 5 and 1 (α = 0.10, 0.05 and 0.01). The average VaRs reported here are the observed average 1% quantiles of the density forecasts. The coverages correspond with the observed fraction of returns below the respective VaRs. The average ES values are equal to the conditional mean return, given a realization below the predicted VaR. Backtesting methods for VaR are labeled as UC, IND and CC. Here, we provide p-values for these tests. The last two rows for each block report McNeil-Frey test statistics and corresponding p-values for backtesting the ES estimates.

		C-Sket	E-Norm	C-Lap	E-Norm * C-Lap	5-GARCH	5-EGARCH	5-TGARCH	5-CGARCH	4-Normal	4-Student t	4-GED	4-Laplace	4-Skewed t	Equal	20Full
CON	Av. VaR	-0.0256	-0.0216	-0.0287	-0.0247	-0.0249	-0.0247	-0.0246	-0.0274	-0.0239	-0.0243	-0.0241	-0.0276	-0.0248	-0.0252	-0.0256
	Coverage	0.0080	0.0160	0.0046	0.0104	0.0092	0.01	0.0097	0.0063	0.0143	0.0105	0.0112	0.005	0.0095	0.0075	0.0085
	UC(p)	0.343	0.010	0.005	0.839	0.719	0.985	0.896	0.068	0.06	0.985	0.595	0.012	0.806	0.233	0.567
	IND(p)	1.000	0.557	1.000	0.612	0.522	0.576	0.558	0.318	0.442	0.576	1.000	1.000	1.000	1.000	1.000
	CC(p)	0.637	0.032	0.021	0.861	0.764	0.855	0.835	0.115	0.128	0.855	0.868	0.044	0.97	0.491	0.849
	Av. ES	-0.0297	-0.0255	-0.0373	-0.0277	-0.0293	-0.0285	-0.0284	-0.0331	-0.0278	-0.0342	-0.0300	-0.0353	-0.0282	-0.0324	-0.0321
	M-F	0.9611	-0.3329	3.0859	-0.3482	-0.8413	-0.5963	-0.5671	-1.7341	-0.3708	-2.0785	-1.0762	-2.4038	-0.4973	-1.5528	-1.4951
	M-F(p)	0.3416	0.7392	0.0021	0.7277	0.4002	0.5510	0.5707	0.0829	0.7108	0.0377	0.2819	0.0163	0.6190	0.1205	0.1349
CEN	Av. VaR	-0.0256	-0.0216	-0.0287	-0.0239	-0.0248	-0.0246	-0.0245	-0.0276	-0.222	-0.0244	-0.0236	-0.0277	-0.0248	-0.0252	-0.0238
	Coverage	0.0080	0.0160	0.0046	0.0121	0.0095	0.0095	0.0102	0.0058	0.0151	0.0106	0.0138	0.0053	0.0097	0.0075	0.0104
	UC(p)	0.343	0.010	0.005	0.338	0.806	0.806	0.926	0.037	0.03	0.985	0.093	0.018	0.895	0.233	0.869
	IND(p)	1.000	0.557	1.000	0.310	0.540	1.000	0.594	1.000	0.49	0.576	1.000	1.000	0.558	1.000	0.318
	CC(p)	0.637	0.032	0.021	0.337	0.804	0.97	0.864	0.113	0.074	0.855	0.245	0.062	0.835	0.491	0.600
	Av. ES	-0.0297	-0.0255	-0.0373	-0.0262	-0.0288	-0.0283	-0.0279	-0.0352	-0.0262	-0.0341	-0.0273	-0.0344	-0.0277	-0.0324	-0.0289
	M-F	0.9611	-0.3329	3.0859	0.1132	-0.6899	-0.5289	-0.4018	-2.3791	0.1085	-2.0417	-0.2103	-2.1321	-0.3315	-1.5528	-0.7155
	M-F(p)	0.3416	0.7392	0.0021	0.9099	0.4903	0.5969	0.6878	0.0174	0.9136	0.0412	0.8334	0.0330	0.7403	0.1205	0.4743

Figure 6: VaR forecasts

This figure presents the $S\&P500$ log-returns for the period Oktober 25, 1991 - March 14, 2008 and out-of-sample 95% and 99% VaR forecasts derived from the combined pool of 20 models using logarithmic, cl and csl scoring rules.





5 Conclusion

Managing uncertainty in financial decision making and financial risk management is an important concern in many applications. During the last decades, much relevance is imposed on this topic where predictive densities have received increasing attention in economics and finance because a density forecast provides information on the uncertainties associated with the forecast. In the financial industry, many practitioners have developed their own risk management system and they obtain VaR estimates for their asset portfolios commonly from a single (predictive density) model. However, relying upon a single model for VaR estimation is dangerous. This leads us to consider combinations of predictive density similar to that of portfolio optimization allowing the possibility that all of the models under consideration are false. In combining predictive densities, the way how we decide to measure the accuracy of the resulting mixture is essential in determining the construction of the optimal pool. We have shown that both conditional likelihood and censored likelihood scoring rules are convenient metrics in comparing density forecasts when interest lies in a region instead of the whole distribution. The underlying idea behind *cl* and *csl* scoring rules is that they replace the full likelihood by the conditional likelihood, given that the actual observation lies in the region of interest, or by the censored likelihood, with censoring of the observations outside the region of interest.

In our study, we have proposed a methodology that aims to combine density forecasts by selecting the optimal weights based on these scoring rules. Furthermore, we have applied methods from different perspectives in comparing the prediction models and we have confirmed the link between these methods. For instance, a higher score based on suitable scoring rules indicates more accurate VaR and ES estimates. In our study, we aim to improve VaR and ES estimates by developing approaches that result in higher predictive density scores. We have considered 4 types of volatility models and each has the choice of 5 different distributions of the innovations, allowing us to create 20 prediction models and compared them by different scoring criteria. It is essential to mention that we are aware of the fact that there the models included in our collection are far from being complete. Many distributions that have been proven in the literature are not discussed. From the results of the individual prediction models, we have observed that none of the individual models such as Student t and Skewed t type models outperform others consistently in terms of predictive power according to *cl* and *csl* scoring rules, signifying the danger of putting all the weights on one single model. Furthermore, we have confirmed that in density forecast the choice of conditional distribution is more important than the choice of conditional volatility models.

The outcome of several analysis of the individual predictive models suggested that there is room for improvement in predictive density accuracy. We have applied the combining methodology proposed in his study first ex post. Moreover, we first assume that we have access of the entire data up to the end of the outsample at a point of time of the sample period. This scheme could not be used in practice since only past data are available. However, this setting is illustrative as a starting point in the large area of combining density models before we consider real time combining. Based on this strategy, we have found evidence that there indeed exists a optimal combination of models that is capable in outperforming even the best individual models. One of the most important finding from combining densities is that including relatively poor performers in pools of multiple models could lead to more accurate forecasts which underlines the crucial intuition behind combining.

A great issue in combining is that the number of combining possibilities increases greatly when the number of models included increases. It is time consuming to consider all combinations. Therefore, we have developed a strategy that could "guess" the most influential models in the optimal pool without checking all combinations. We have noticed that a certain model can be assigned as a big contributor in the optimal pool when this model is both dominant in its volatility model family and its distribution family. The result of the optimal pool consisting all models confirmed this strategy. We have also observed that our optimal pool outperforms the best individual models. The score differences however decrease for smaller regions of interests due to the smaller number of observations within these areas. Furthermore, we advocate for combining using different models because of diversification as even relatively poor performers in pools of multiple models could lead to more accurate forecasts. Finally, by performing several back tests on the resulting VaR and ES estimates of the combined density models, we have demonstrated that the accuracy of these estimates are considerable improved. A comparison in accuracy between the applied scoring rules for computing the density forecasts shows that VaR and ES estimates are more accurate through combining density models using conditional and censored likelihood scoring rules than logarithmic scoring rule. From the outcome of this research, it seems to us that this is a lesson that should be extended in the future of financial decision making. It is an avenue open for further investigation to develop even more advanced combining techniques in predictive densities.

Acknowledgements

I would like to thank my supervisor Prof. dr. Dick van Dijk for his continues support, for his detailed and concerned attitude towards the evaluation and feedback of my thesis. I am especially grateful for his knowledge which has helped me from the initial to the final level to develop an understanding of the subject.

I would also like to acknowledge dr. Erik Kole as the second reader of this thesis, and I am gratefully indebted to him for his very valuable comments on this thesis.

A Other figures and tables

TABLE 15: PREDICTIVE DENSITY EVALUATION FOR COMBINED MODELS - EX POST OPTIMAL CONDITIONAL LIKELIHOOD PREDICTIVE SCORES FOR $\alpha = 0.10$

In this table we presents the conditional likelihood predictive scores of the density forecasts for pools of two models. The scores presented in this table are the sum of scores over the evaluation period for the regional distribution $\alpha = 0.10$, evaluated over the out-of-sample period October 25, 1991 - March 14, 2008 (4117 observations). Entries above the diagonal are log scores of optimal pools. The corresponding weights of the pool for the model in that row are presented in entries below the diagonal. Note that the numbers in the first row and column correspond with the prediction models similar to previous tables, where each model is assigned with a number. The ”target score” is equal to -302, referring to the best individual performance achieved by CGARCH - Student *t*. This target score is surpassed by 42 combinations of two models, as highlighted in grey. The best performing pool of two models consists EGARCH - Normal and CGARCH - Laplace, which has achieved a score equals to -297

	G-Norm	G-Stu*t*	G-GED	G-Lap	G-Ske*t*	E-Norm	E-Stu*t*	E-GED	E-Lap	E-Ske*t*	T-Norm	T-Stu*t*	T-GED	T-Lap	T-Ske*t*	C-Norm	C-Stu*t*	C-GED	C-Lap	C-Ske*t*
G-Norm		-304	-319	-304	-304	-357	-305	-312	-305	-304	-372	-303	-323	-303	-303	-362	-302	-323	-301	-302
G-Stu*t*	1.00		-303	-303	-303	-302	-304	-299	-303	-303	-304	-302	-302	-302	-302	-303	-302	-302	-302	-302
G-GED	1.00	0.23		-302	-303	-318	-304	-311	-303	-303	-319	-303	-318	-302	-302	-309	-302	-308	-301	-301
G-Lap	0.63	0.27	0.42		-303	-300	-303	-297	-307	-303	-302	-301	-299	-306	-302	-306	-302	-305	-307	-302
G-Ske*t*	1.00	0.51	0.67	0.83		-302	-303	-299	-304	-303	-303	-302	-301	-303	-302	-303	-302	-303	-303	-302
E-Norm	0.74	0.27	0.24	0.49	0.32		-303	-312	-301	-302	-361	-303	-322	-300	-302	-346	-300	-319	**-297**	-300
E-Stu*t*	1.00	0.24	0.67	0.66	0.35	0.74		-300	-304	-304	-304	-302	-302	-302	-302	-304	-302	-303	-302	-302
E-GED	1.00	0.61	0.85	0.72	0.63	1.00	0.63		-298	-299	-312	-301	-312	-298	-300	-303	-298	-302	-296	-298
E-Lap	0.65	0.23	0.41	0.30	0.13	0.54	0.26	0.27		-304	-304	-301	-300	-306	-302	-307	-302	-306	-307	-302
E-Ske*t*	0.96	0.35	0.62	0.72	0.32	0.70	0.70	0.36	0.86		-304	-302	-301	-303	-302	-304	-302	-303	-303	-302
T-Norm	0.45	0.05	0.01	0.39	0.13	0.00	0.09	0.00	0.37	0.14		-303	-323	-303	-303	-361	-302	-323	-300	-301
T-Stu*t*	1.00	0.68	0.87	0.74	0.66	0.90	0.71	0.51	0.77	0.69	1.00		-302	-301	-303	-301	-301	-300	-299	-301
T-GED	1.00	0.46	0.42	0.62	0.50	0.80	0.49	0.00	0.62	0.51	1.00	0.24		-300	-302	-309	-300	-308	-297	-299
T-Lap	0.64	0.32	0.43	0.75	0.28	0.53	0.38	0.31	0.73	0.35	0.63	0.26	0.40		-302	-305	-301	-305	-306	-302
T-Ske*t*	1.00	0.64	0.76	0.78	0.66	0.80	0.68	0.46	0.80	0.68	1.00	0.43	0.64	0.78		-302	-301	-301	-301	-301
C-Norm	0.14	0.05	0.08	0.19	0.05	0.16	0.05	0.07	0.16	0.04	0.21	0.07	0.09	0.19	0.08		-302	-328	-306	-302
C-Stu*t*	0.95	1.00	0.75	0.86	1.00	0.69	1.00	0.40	0.94	1.00	0.84	0.51	0.55	0.77	0.57	0.94		-301	-301	-302
C-GED	0.54	0.14	0.15	0.30	0.15	0.47	0.16	0.09	0.29	0.16	0.56	0.13	0.15	0.28	0.15	0.76	0.02		-305	-302
C-Lap	0.54	0.29	0.36	0.21	0.25	0.46	0.33	0.27	0.38	0.30	0.53	0.28	0.35	0.25	0.26	0.72	0.07	0.95		-302
C-Ske*t*	0.86	0.71	0.65	1.00	0.94	0.64	0.77	0.38	1.00	0.93	0.78	0.46	0.51	0.84	0.50	0.94	0.31	0.86	0.91	

Table 16: Predictive Density Evaluation for combined models - Ex post optimal censored likelihood predictive scores for $\alpha = 0.10$

In this table we presents the censored likelihood predictive scores of the density forecasts for pools of two models. The scores presented in this table are the sum of scores over the evaluation period for the regional distribution $\alpha = 0.10$, evaluated over the out-of-sample period October 25, 1991 - March 14, 2008 (4117 observations). Entries above the diagonal are log scores of optimal pools. The corresponding weights of the pool for the model in that row are presented in entries below the diagonal. Note that the numbers in the first row and column correspond with the prediction models similar to previous tables, where each model is assigned with a number. The "target score" is equal to -1553, referring to the best individual performance achieved by CGARCH - Skewed t. This target score is surpassed by 29 combinations of two models, as highlighted in grey. The best performing pool of two models consists EGARCH - Normal and CGARCH - Laplace, which has achieved a score equals to -1546

	G-Norm	G-Stu*t*	G-GED	G-Lap	G-Ske*t*	E-Norm	E-Stu*t*	E-GED	E-Lap	E-Ske*t*	T-Norm	T-Stu*t*	T-GED	T-Lap	T-Ske*t*	C-Norm	C-Stu*t*	C-GED	C-Lap	C-Ske*t*
G-Norm		-1561	-1596	-1563	-1559	-1606	-1567	-1577	-1566	-1566	-1634	-1556	-1598	-1560	-1554	-1623	-1556	-1596	-1561	-1554
G-Stu*t*	1.00		-1561	-1560	-1559	-1552	-1561	-1556	-1560	-1561	-1560	-1555	-1561	-1557	-1554	-1558	-1556	-1559	-1557	-1554
G-GED	0.65	0.00		-1567	-1559	-1577	-1568	-1579	-1571	-1567	-1590	-1556	-1598	-1563	-1554	-1580	-1556	-1583	-1567	-1554
G-Lap	0.67	0.19	0.82		-1558	-1549	-1563	-1557	-1568	-1562	-1560	-1553	-1565	-1564	-1552	-1567	-1555	-1568	-1567	-1554
G-Ske*t*	1.00	1.00	1.00	0.85		-1551	-1559	-1554	-1559	-1559	-1558	-1554	-1559	-1556	-1554	-1556	-1555	-1557	-1556	-1554
E-Norm	0.83	0.46	0.65	0.59	0.45		-1556	-1573	-1553	-1556	-1611	-1552	-1582	-1550	-1551	-1591	-1548	-1575	**-1546**	-1547
E-Stu*t*	0.78	0.00	0.82	0.54	0.00	0.49		-1560	-1568	-1569	-1564	-1555	-1566	-1560	-1554	-1565	-1556	-1566	-1561	-1554
E-GED	0.80	0.42	0.91	0.63	0.41	0.55	0.54		-1560	-1559	-1577	-1554	-1579	-1557	-1553	-1561	-1553	-1565	-1555	-1551
E-Lap	0.61	0.12	0.65	0.00	0.10	0.40	0.25	0.32		-1567	-1563	-1554	-1568	-1564	-1553	-1572	-1556	-1573	-1569	-1554
E-Ske*t*	0.77	0.15	0.77	0.54	0.00	0.49	0.56	0.46	0.73		-1564	-1555	-1565	-1559	-1554	-1564	-1556	-1565	-1561	-1554
T-Norm	0.54	0.17	0.44	0.41	0.15	0.02	0.30	0.21	0.44	0.31		-1556	-1598	-1560	-1554	-1618	-1554	-1594	-1556	-1552
T-Stu*t*	1.00	0.79	1.00	0.83	0.71	0.68	0.91	0.79	0.88	0.88	1.00		-1556	-1553	-1554	-1551	-1552	-1552	-1550	-1551
T-GED	0.60	0.07	0.41	0.36	0.11	0.36	0.28	0.00	0.44	0.30	0.56	0.00		-1563	-1554	-1578	-1555	-1582	-1563	-1553
T-Lap	0.69	0.34	0.82	0.87	0.31	0.45	0.53	0.44	1.00	0.53	0.66	0.16	0.77		-1552	-1563	-1554	-1564	-1563	-1553
T-Ske*t*	1.00	0.75	1.00	0.82	0.74	0.68	0.87	0.74	0.86	0.90	1.00	0.77	1.00	0.84		-1550	-1550	-1551	-1549	-1550
C-Norm	0.09	0.05	0.14	0.07	0.05	0.10	0.08	0.12	0.11	0.09	0.14	0.06	0.16	0.08	0.06		-1553	-1634	-1570	-1552
C-Stu*t*	0.98	1.00	1.00	0.94	0.78	0.56	1.00	0.65	1.00	1.00	0.81	0.45	0.96	0.77	0.45	0.95		-1554	-1555	-1554
C-GED	0.21	0.03	0.09	0.00	0.03	0.17	0.06	0.07	0.03	0.08	0.23	0.03	0.12	0.00	0.04	0.54	0.03		-1571	-1552
C-Lap	0.53	0.24	0.59	0.27	0.22	0.35	0.38	0.32	0.56	0.39	0.48	0.21	0.50	0.24	0.22	0.87	0.08	1.00		-1553
C-Ske*t*	0.99	1.00	1.00	0.96	1.00	0.57	1.00	0.65	1.00	1.00	0.83	0.50	0.91	0.80	0.48	0.95	1.00	0.97	0.94	

Table 17: Predictive Density Evaluation for combined models - Ex post optimal conditional likelihood predictive scores for $\alpha = 0.05$

In this table we presents the conditional likelihood predictive scores of the density forecasts for pools of two models. The scores presented in this table are the sum of scores over the evaluation period for the regional distribution $\alpha = 0.05$, evaluated over the out-of-sample period October 25, 1991 - March 14, 2008 (4117 observations). Entries above the diagonal are log scores of optimal pools. The corresponding weights of the pool for the model in that row are presented in entries below the diagonal. Note that the numbers in the first row and column correspond with the prediction models similar to previous tables, where each model is assigned with a number. The ”target score” is equal to -165, referring to the best individual performance achieved by CGARCH - Student t. This target score is surpassed by 2 combinations of two models, as highlighted in grey. The best performing pool of two models consists CGARCH - Normal and CGARCH - Student t, which has achieved a score equals to -164

	G-Norm	G-Stu*t*	G-GED	G-Lap	G-Ske*t*	E-Norm	E-Stu*t*	E-GED	E-Lap	E-Ske*t*	T-Norm	T-Stu*t*	T-GED	T-Lap	T-Ske*t*	C-Norm	C-Stu*t*	C-GED	C-Lap	C-Ske*t*
G-Norm		-166	-183	-168	-167	-221	-167	-179	-170	-168	-232	-168	-189	-169	-169	-220	-165	-183	-166	-166
G-Stu*t*	1.00		-166	-166	-166	-166	-166	-166	-166	-166	-166	-166	-166	-166	-166	-165	-165	-165	-166	-166
G-GED	1.00	0.00		-168	-167	-183	-167	-178	-169	-168	-183	-168	-183	-169	-169	-174	-165	-173	-167	-166
G-Lap	0.68	0.06	0.52		-167	-167	-167	-167	-170	-168	-169	-167	-168	-170	-168	-166	-165	-167	-170	-166
G-Ske*t*	1.00	0.00	1.00	1.00		-167	-167	-166	-167	-167	-167	-167	-167	-167	-167	-165	-165	-166	-167	-166
E-Norm	0.69	0.05	0.05	0.38	0.14		-167	-179	-169	-168	-224	-168	-189	-168	-169	-208	-165	-182	-165	-166
E-Stu*t*	1.00	0.00	1.00	0.79	0.30	0.93		-167	-167	-167	-167	-167	-167	-167	-167	-166	-165	-166	-166	-166
E-GED	1.00	0.27	0.79	0.61	0.38	1.00	0.35		-167	-167	-179	-168	-179	-167	-168	-171	-165	-171	-165	-165
E-Lap	0.70	0.00	0.51	0.00	0.00	0.64	0.03	0.38		-168	-170	-167	-169	-171	-168	-167	-165	-168	-170	-166
E-Ske*t*	0.98	0.00	0.92	0.82	0.00	0.85	0.00	0.57	1.00		-168	-167	-168	-168	-168	-166	-165	-167	-167	-166
T-Norm	0.34	0.00	0.00	0.28	0.00	0.00	0.00	0.00	0.26	0.00		-168	-189	-170	-169	-221	-165	-183	-167	-166
T-Stu*t*	1.00	0.00	1.00	0.75	0.31	1.00	0.40	0.82	0.81	0.56	1.00		-168	-167	-168	-166	-165	-166	-166	-166
T-GED	1.00	0.00	0.12	0.49	0.13	0.88	0.06	0.00	0.50	0.21	1.00	0.00		-169	-169	-176	-165	-175	-166	-166
T-Lap	0.65	0.08	0.47	0.00	0.00	0.61	0.19	0.39	0.42	0.16	0.70	0.19	0.49		-168	-167	-165	-168	-170	-166
T-Ske*t*	0.97	0.00	0.89	0.73	0.00	0.89	0.19	0.64	0.80	0.32	1.00	0.00	0.91	0.83		-166	-165	-167	-167	-166
C-Norm	0.30	0.19	0.20	0.35	0.22	0.30	0.20	0.19	0.34	0.22	0.37	0.22	0.20	0.37	0.25		**-164**	-182	-164	-164
C-Stu*t*	1.00	1.00	1.00	1.00	1.00	0.87	1.00	0.74	1.00	1.00	1.00	1.00	0.98	1.00	1.00	0.78		-165	-165	-165
C-GED	0.80	0.22	0.31	0.47	0.27	0.74	0.24	0.19	0.46	0.28	0.87	0.25	0.32	0.48	0.31	0.78	0.23		-166	-165
C-Lap	0.61	0.22	0.48	0.68	0.21	0.57	0.31	0.38	0.81	0.31	0.65	0.32	0.48	0.83	0.36	0.61	0.07	0.51		-166
C-Ske*t*	0.89	0.53	0.94	1.00	1.00	0.79	0.86	0.62	1.00	1.00	0.92	0.75	0.81	1.00	1.00	0.75	0.00	0.73	1.00	

Table 18: Predictive Density Evaluation for combined models - Ex post optimal censored likelihood predictive scores for $\alpha = 0.05$

In this table we presents the censored likelihood predictive scores of the density forecasts for pools of two models. The scores presented in this table are the sum of scores over the evaluation period for the regional distribution $\alpha = 0.05$, evaluated over the out-of-sample period October 25, 1991 - March 14, 2008 (4117 observations). Entries above the diagonal are log scores of optimal pools. The corresponding weights of the pool for the model in that row are presented in entries below the diagonal. Note that the numbers in the first row and column correspond with the prediction models similar to previous tables, where each model is assigned with a number. The ”target score” is equal to -966, referring to the best individual performance achieved by TGARCH - Skewed *t*. This target score is surpassed by 35 combinations of two models, as highlighted in grey. The best performing pool of two models consists EGARCH - Normal and CGARCH - Laplace, which has achieved a score equals to -955

	G-Norm	G-Stu*t*	G-GED	G-Lap	G-Ske*t*	E-Norm	E-Stu*t*	E-GED	E-Lap	E-Ske*t*	T-Norm	T-Stu*t*	T-GED	T-Lap	T-Ske*t*	C-Norm	C-Stu*t*	C-GED	C-Lap	C-Ske*t*
G-Norm		-972	-999	-972	-971	-1011	-974	-982	-974	-974	-1034	-966	-1000	-969	-966	-1025	-967	-996	-969	-967
G-Stu*t*	0.94		-972	-971	-971	-962	-972	-966	-971	-972	-969	-966	-971	-968	-965	-969	-968	-969	-969	-968
G-GED	0.69	0.00		-976	-971	-982	-975	-983	-977	-974	-993	-966	-997	-972	-965	-984	-968	-986	-975	-968
G-Lap	0.58	0.18	0.67		-970	-959	-972	-965	-977	-972	-967	-964	-971	-973	-963	-976	-968	-977	-976	-968
G-Ske*t*	0.96	0.74	0.98	0.84		-962	-971	-965	-971	-971	-969	-965	-969	-968	-965	-969	-968	-969	-968	-968
E-Norm	0.82	0.53	0.66	0.65	0.54		-965	-979	-962	-965	-1015	-962	-987	-959	-962	-997	-958	-980	**-955**	-958
E-Stu*t*	0.78	0.11	0.81	0.63	0.17	0.46		-967	-975	-975	-971	-966	-972	-969	-965	-973	-968	-973	-971	-968
E-GED	0.85	0.54	0.92	0.72	0.54	0.57	0.59		-967	-966	-982	-964	-983	-965	-963	-969	-963	-971	-963	-962
E-Lap	0.56	0.13	0.57	0.01	0.12	0.36	0.18	0.26		-975	-969	-964	-973	-973	-964	-979	-968	-980	-978	-968
E-Ske*t*	0.78	0.20	0.76	0.63	0.11	0.44	0.61	0.40	0.84		-971	-966	-971	-970	-965	-973	-968	-973	-971	-968
T-Norm	0.59	0.31	0.47	0.51	0.31	0.03	0.36	0.22	0.52	0.37		-966	-998	-967	-966	-1019	-964	-993	-963	-964
T-Stu*t*	1.00	0.87	1.00	0.86	0.83	0.61	0.90	0.71	0.88	0.89	1.00		-966	-964	-965	-962	-963	-962	-961	-963
T-GED	0.68	0.25	0.56	0.54	0.30	0.39	0.38	0.00	0.57	0.40	0.59	0.00		-970	-965	-981	-967	-983	-969	-966
T-Lap	0.63	0.38	0.74	1.00	0.37	0.39	0.50	0.36	1.00	0.52	0.56	0.15	0.60		-964	-972	-966	-973	-972	-967
T-Ske*t*	1.00	0.80	0.96	0.85	0.84	0.60	0.83	0.65	0.87	0.89	1.00	0.58	0.94	0.85		-962	-963	-962	-961	-963
C-Norm	0.08	0.05	0.11	0.08	0.05	0.09	0.06	0.08	0.10	0.05	0.13	0.05	0.12	0.07	0.05		-966	-1022	-979	-966
C-Stu*t*	0.87	1.00	1.00	0.95	0.86	0.47	1.00	0.51	1.00	1.00	0.66	0.37	0.78	0.73	0.40	0.95		-966	-967	-967
C-GED	0.23	0.04	0.06	0.02	0.04	0.19	0.05	0.04	0.04	0.06	0.25	0.04	0.08	0.01	0.04	0.58	0.03		-981	-967
C-Lap	0.44	0.22	0.44	0.19	0.21	0.30	0.30	0.24	0.42	0.30	0.39	0.19	0.36	0.16	0.19	0.81	0.08	0.98		-968
C-Ske*t*	0.84	0.80	0.97	0.96	0.89	0.45	0.98	0.48	1.00	1.00	0.64	0.35	0.69	0.72	0.35	0.95	0.42	0.96	0.93	

Table 19: Predictive Density Evaluation - Real Time - Average weights -

This table presents the average weight through recursively updating of optimal pool using only the data available on each date. Rows in each block display the individual models forming that pool reported in Table 12 of combined models. Note that some model weights are highlighted, meaning that these models are both dominant in their volatility model family and distribution family.

		Log predictive Scores	Conditional Likelihood Scores			Censored Likelihood Scores		
		Whole distribution	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
5 GARCH Models	GARCH - Normal	0.24	0.16	0.24	0.15	0.22	0.16	0.24
	GARCH - Student t	0.25	0.24	0.25	0.47	0.07	0.24	0.40
	GARCH - GED	0.12	0.18	0.12	0.02	0.11	0.16	0.12
	GARCH - Laplace	0.12	0.40	0.12	0.12	0.34	0.38	0.24
	GARCH - Skewed t	0.27	0.03	0.27	0.24	0.26	0.05	0.00
5 EGARCH Models	EGARCH - Normal	0.41	0.29	0.17	0.10	0.34	0.36	0.37
	EGARCH - Student t	0.13	0.15	0.30	0.41	0.08	0.19	0.19
	EGARCH - GED	0.22	0.23	0.14	0.15	0.22	0.29	0.31
	EGARCH - Laplace	0.12	0.27	0.33	0.10	0.20	0.16	0.13
	EGARCH - Skewed t	0.11	0.07	0.06	0.25	0.16	0.00	0.00
5 TGARCH Models	TGARCH - Normal	0.28	0.14	0.08	0.08	0.24	0.17	0.24
	TGARCH - Student t	0.25	0.22	0.41	0.46	0.06	0.19	0.32
	TGARCH - GED	0.10	0.22	0.14	0.14	0.10	0.15	0.12
	TGARCH - Laplace	0.10	0.40	0.36	0.15	0.33	0.35	0.30
	TGARCH - Skewed t	0.26	0.01	0.01	0.17	0.27	0.13	0.01
5 CGARCH Models	CGARCH - Normal	0.07	0.12	0.23	0.19	0.12	0.12	0.26
	CGARCH - Student t	0.45	0.53	0.36	0.41	0.29	0.58	0.61
	CGARCH - GED	0.00	0.03	0.07	0.10	0.00	0.00	0.00
	CGARCH - Laplace	0.11	0.29	0.32	0.21	0.16	0.15	0.11
	CGARCH - Skewed t	0.37	0.03	0.02	0.09	0.42	0.15	0.00
4 Normal Models	GARCH - Normal	0.09	0.24	0.25	0.13	0.10	0.15	0.11
	EGARCH - Normal	0.71	0.50	0.47	0.46	0.64	0.59	0.57
	TGARCH - Normal	0.02	0.15	0.01	0.02	0.03	0.06	0.05
	CGARCH - Normal	0.19	0.11	0.26	0.19	0.23	0.19	0.26
4 Student t Models	GARCH - Student t	0.06	0.06	0.10	0.09	0.08	0.06	0.00
	EGARCH - Student t	0.04	0.09	0.14	0.12	0.00	0.01	0.23
	TGARCH - Student t	0.46	0.47	0.20	0.12	0.44	0.53	0.53
	CGARCH - Student t	0.44	0.39	0.57	0.67	0.48	0.40	0.24
4 GED Models	GARCH - GED	0.09	0.12	0.20	0.17	0.07	0.07	0.17
	EGARCH - GED	0.84	0.48	0.50	0.43	0.81	0.66	0.55
	TGARCH - GED	0.07	0.32	0.04	0.03	0.11	0.25	0.20
	CGARCH - GED	0.01	0.09	0.27	0.36	0.09	0.01	0.08
4 Laplace Models	GARCH - Laplace	0.17	0.02	0.12	0.03	0.23	0.15	0.04
	EGARCH - Laplace	0.09	0.19	0.29	0.20	0.09	0.11	0.23
	TGARCH - Laplace	0.41	0.31	0.12	0.10	0.43	0.51	0.48
	CGARCH - Laplace	0.34	0.48	0.47	0.66	0.24	0.24	0.26
4 Skewed t Models	GARCH - Skewed t	0.09	0.04	0.12	0.03	0.10	0.06	0.01
	EGARCH - Skewed t	0.03	0.09	0.15	0.22	0.00	0.02	0.23
	TGARCH - Skewed t	0.40	0.48	0.16	0.08	0.44	0.54	0.55
	CGARCH - Skewed t	0.48	0.38	0.57	0.66	0.46	0.39	0.21

Table 20: Predictive Density Evaluation - Real Time - Average weights - Pool of full collection of models

	Log predictive Scores	Conditional Likelihood Scores			Censored Likelihood Scores		
	Whole distribution	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
GARCH - Normal	0.00	0.00	0.00	0.00	0.00	0.00	0.05
GARCH - Student t	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GARCH - GED	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GARCH - Laplace	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GARCH - Skewed t	0.05	0.00	0.00	0.00	0.00	0.00	0.00
EGARCH - Normal	0.43	0.25	0.11	0.00	0.35	0.40	0.30
EGARCH - Student t	0.00	0.00	0.00	0.00	0.00	0.00	0.00
EGARCH - GED	0.13	0.10	0.10	0.10	0.14	0.10	0.17
EGARCH - Laplace	0.00	0.00	0.15	0.00	0.00	0.00	0.05
EGARCH - Skewed t	0.00	0.00	0.00	0.00	0.00	0.00	0.00
TGARCH - Normal	0.00	0.00	0.00	0.00	0.00	0.00	0.00
TGARCH - Student t	0.00	0.15	0.00	0.00	0.00	0.12	0.00
TGARCH - GED	0.00	0.15	0.00	0.00	0.07	0.12	0.05
TGARCH - Laplace	0.00	0.00	0.00	0.00	0.00	0.00	0.00
TGARCH - Skewed t	0.00	0.00	0.00	0.00	0.05	0.00	0.00
CGARCH - Normal	0.00	0.00	0.13	0.15	0.00	0.00	0.10
CGARCH - Student t	0.17	0.00	0.22	0.38	0.05	0.10	0.10
CGARCH - GED	0.00	0.00	0.06	0.07	0.00	0.00	0.00
CGARCH - Laplace	0.07	0.34	0.20	0.19	0.12	0.16	0.17
CGARCH - Skewed t	0.15	0.00	0.02	0.10	0.22	0.00	0.00
Score	-5184	-298	-163	-42	-1546	-961	-301

Table 21: Predictive Density Evaluation - Real Time - VaR and ES characteristics - Alpha = 0.10

This table summarizes the VaR and ES as risk measures with several additional backtesting approaches for each prediction model presented in columns. The rows are separated in three blocks, where each block corresponds with a region of interest given by quantiles. We consider three quantiles, 10, 5 and 1 (α = 0.10, 0.05 and 0.01). The average VaRs reported here are the observed average 10% quantiles of the density forecasts. The coverages correspond with the observed fraction of returns below the respective VaRs. The average ES values are equal to the conditional mean return, given a realization below the predicted VaR. Backtesting methods for VaR are labeled as UC, IND and CC. Here, we provide p-values for these tests. The last two rows for each block report McNeil-Frey test statistics and corresponding p-values for backtesting the ES estimates.

		C-Ske$t$	E-Norm	C-Lap	E-Norm * C-Lap	5-GARCH	5-EGARCH	5-TGARCH	5-CGARCH	4-Normal	4-Student t	4-GED	4-Laplace	4-Skewed t	Equal	20Full
CON	Av. VaR	-0.0113	-0.0117	-0.0118	-0.0115	-0.0113	-0.0114	-0.0121	-0.0122	-0.0119	-0.0109	-0.0109	-0.0114	-0.0111	-0.0134	-0.0116
	Coverage	0.1018	0.0896	0.0908	0.0964	0.0993	0.0998	0.1001	0.0889	0.0925	0.0989	0.1064	0.0984	0.1066	0.0974	0.0959
	UC(p)	0.787	0.097	0.155	0.582	0.920	0.979	0.991	0.084	0.248	0.861	0.332	0.803	0.314	0.689	0.614
	IND(p)	0.059	0.244	0.158	0.153	0.376	0.157	0.246	0.171	0.542	0.129	0.341	0.072	0.057	0.185	0.360
	CC(p)	0.163	0.138	0.134	0.310	0.672	0.367	0.510	0.088	0.426	0.310	0.397	0.192	0.098	0.384	0.579
	Av. ES	-0.0167	-0.0166	-0.0178	-0.0174	-0.0175	0.01768	-0.0174	-0.0185	-0.0177	-0.0183	0.0171	-0.0174	-0.0164	-0.0178	-0.0176
	M-F	-1.1952	-1.3465	2.4432	1.2351	1.5315	2.1133	1.0944	4.2183	2.1038	3.6132	0.2137	1.0763	-2.1057	2.4212	1.8136
	M-F(p)	0.2321	0.1782	0.0146	0.2169	0.1257	0.0346	0.2738	0.0000	0.0354	0.0003	0.8308	0.2818	0.035	0.0155	0.0698
CEN	Av. VaR	-0.0113	-0.0117	-0.0118	-0.0115	-0.0107	-0.0109	-0.011	-0.0121	-0.0119	-0.011	-0.0109	-0.0114	-0.0111	-0.0134	-0.0113
	Coverage	0.1018	0.0896	0.0908	0.0969	0.1025	0.1042	0.1013	0.0887	0.093	0.0989	0.1076	0.0959	0.1074	0.0974	0.1003
	UC(p)	0.787	0.097	0.155	0.635	0.703	0.522	0.844	0.077	0.281	0.861	0.249	0.532	0.265	0.689	0.969
	IND(p)	0.059	0.244	0.158	0.135	0.576	0.285	0.207	0.077	0.349	0.129	0.547	0.067	0.107	0.185	0.227
	CC(p)	0.163	0.138	0.134	0.293	0.795	0.460	0.442	0.044	0.360	0.310	0.430	0.154	0.146	0.384	0.482
	Av. ES	-0.0167	-0.0166	-0.0178	-0.0173	-0.0172	-0.0172	-0.0171	-0.0184	-0.0176	-0.0183	-0.0169	-0.0175	-0.0164	-0.0178	-0.0171
	M-F	-1.1952	-1.3465	2.4432	0.9465	0.6514	0.6471	0.2315	3.9111	1.8018	3.5483	-0.5511	1.4982	-2.1133	2.4212	0.2046
	M-F(p)	0.2321	0.1782	0.0146	0.3439	0.5148	0.5176	0.8169	0.0000	0.0716	0.0003	0.5816	0.1341	0.0346	0.0155	0.8379

References

K. Aas and I.H. Haff. The generalized hyperbolic skew student's t-distribution. *Journal of Financial Econometrics*, 4(2), 275-309, 2006.

G. Amisano and R. Giacomini. Comparing density forecasts via weighted likelihood ratio tests. *American Statistical Association*, 25, 177-190, 2007.

Y. Bao, T. Lee, and B. Saltglu. Comparing density forecasts models. *Journal of Forecasting*, 26, 203-225, 2004.

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

T. Bollerslev. Generalised autoregressive conditional heteroscedasticity. *Journal of Econometrics*, 31, 307-327, 1986.

P. Christoffersen. Evaluating interval forecasts. *International Economic Review*, 39, 841-862, 1998.

V. Corradi and N.R. Swanson. Predictive density evaluation. *Handbook of Economic Forecasting*, 1, 197-284, 2006.

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

C. Diks, V. Pancheko, and D. Van Dijk. Likelihood-based scoring rules for comparing density forecasts in tails. *Journal of Econometrics*, 163, 215-230, 2011.

R.F. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of uk inflation. *Econometrica*, 50, 987-1007, 1982.

R.F. Engle and G. Lee. A permanent and transitory component model of stock return volatility. *Cointegration, causality, and forecasting*, 475-497, 1999.

A. Garratt, K. Lee, M.H. Pesaran, and Y. Shin. Forecast uncertainties in macroeconomic modelling: an application to the uk economy. *Journal of the American statistical Assocation*, 98, 829-838, 2003.

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

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

T. Gneiting and R. Ranjan. Comparing density forecasts using threshold and quantile weithed scoring rules. *Technical Report*, 533, 2008.

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

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

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

P.R. Hansen, A. Lunde, and J.M. Nason. The model confidence set for forecasting models. *Federal Reserve Bank of Atlanta Working Paper*, 7, 2005.

A. McNeil and R. Frey. Estimation of tail-related risk measures for heteroscedastic financial time series: An extreme value approach. *Journal of Empirical Finance*, 7, 271-300, 2000.

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

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

A. Timmermann. *Handbook of economic forecasting.* 2006.

K.F. Wallis. Combining density and interval forecasts: A modest proposal. *Oxford Bulletin of Economics and Statistics*, 67, 983-994, 2005.

D.M. Zhu and W. Galbraith. A generalized asymmetric student-t distribution with application to financial econometrics. *Journal of Econometrics*, 157, 297-305, 2006.