Working Paper
Series

FINRISK

National Centre of Competence in Research
Financial Valuation and Risk Management

_____

National Centre of Competence in Research
Financial Valuation and Risk Management

Working Paper No. 689

Weighted Maximum Likelihood for Risk Prediction

Sven C. Steude

_____

FNSNF
SCHWEIZERISCHER NATIONALFONDS
FONDS NATIONAL SUISSE
SWISS NATIONAL SCIENCE FOUNDATION

Weighted Maximum Likelihood for Risk Prediction

Sven C. Steude[a] [*]

[a]*Department of Banking and Finance, University of Zurich, Switzerland*

June 2011

Abstract

Most time series models used in econometrics and empirical finance are estimated with maximum likelihood methods, in particular when interest centers on density and Value–at–Risk (VaR) prediction. The standard maximum likelihood principle implicitly places equal weight on each of the observations in the sample, but depending on the extent to which the model and the true data generating process deviate this can be improved upon. For example, in the context of modeling financial time series, weighting schemes which place relatively more weight on observations in the recent past result in improvement of out–of–sample density forecasts, compared to the default of equal weights. Also, if instead of accurate forecasting of the entire density, interest is restricted to just downside risk, placing more weight on the negative observations in the sample improves results further. In this paper, a third and quite general strategy of shifting more weight towards certain observations of the sample is proposed. Weights are derived from external variables that convey additional information about the true DGP, like trading volume, news arrivals or even investor sentiment. As such, those observations are down weighted that bear a high probability of being destructive outliers with no benefit of using them when fitting the model. Considerable improvements in forecast accuracy for a variety of data sets and different time series models can be realized.

JEL classification: C22; C51; G10

Keywords—GARCH, Value–at–Risk Prediction, Maximum Likelihood, News Arrivals, Trading Volume, News Sentiment.

1 Introduction

In any quantitative discipline time series analysis enjoys itself as an important tool in particular for the applied empiricist. Many of those time series models are estimated with maximum likelihood methods and in fact, in econometrics and empirical finance when interest centers on density or Value-at-Risk (VaR) prediction, likelihood based inference with emphasis on the maximum likelihood estimator (MLE), is at least among if not the, most popular and important.

The original form of the MLE implicitly places the exact same weight on all data inputs. However, similar to the well-known use of weighted least squares to account for discrepancies in the variance of the dependent variable in a linear regression model, weighted likelihood is an established method for estimation and hypothesis testing when the elements of the sample have differing amounts of information. Typically, this is used in the context of robust estimation, to avoid the destructive effect of outliers; see, e.g. Hadi and Luceno (1997) or Cheng (2005). In a related strand of literature, weighted likelihood methodology is used in the area of mixture distributions, see, e.g., Markatou (2000), who uses weighting functions to downweight certain observations with large residuals.

In a pure time series context a first generalization of the traditional MLE is to apply a weighting function such that more recent observations in the sample receive a higher relative weight and values further in the past are down weighted. This idea arises quite naturally and is not new to the literature. Boudoukh, Richardson and Whitelaw (1998) apply it to the simple nonparametric method of historical simulation for example, and Mittnik and Paolella (2000) use a parametric model in conjunction with the weighting scheme.

The method of time-based weighted likelihood (in short, TiWML), is particularly appealing for dealing with the quite nonstandard properties of asset returns, including nonlinear dynamics which change through time, as well as high excess kurtosis and possible time-varying skewness. The most notable stylized fact of asset returns measured at a daily, or higher frequency, is volatility clustering. It is well known that recent volatility shocks are good predictors of volatility in subsequent periods, which would suggest that placing more weight on recent observations could enhance predictive power. Also, for a model whose parameters are changing over time or—more likely in the context of modelling financial returns data—for a much more complicated data generating process which cannot easily be embodied in either a parametric or nonparametric setup, the method of choosing a tractable parametric structure which reasonably captures the salient features of the data generating process and estimating it with more weight given to recent observations can lead to considerable forecasting improvements. This was demonstrated to be the case by Mittnik and Paolella (2000) in the context of density forecasting for financial returns; see, e.g. Diebold, Gunther and Tay (1998), Tay and Wallis (2000), Bao, Lee and Saltoglu (2006), Amisano and Giacomini (2006), and the references therein for further issues related to density forecasting. A notable feature of TiWML is that parameter estimation can be conducted using all relevant, available past data, instead of just an arbitrarily-chosen amount, typically the round numbers of one year (250 trading days) or four years (1000 days). This allows the risk analyst to avoid having to make the difficult decision at what precise point in the past the data are of absolutely no relevance to the risk study (and then implicitly equally weighting all the remaining data).

As is show in Paolella and Steude (2007) a second generalization forms likelihood weights according to their associated location within the residual density and places relatively more

weight on negative (or positive) returns. Within the area of financial risk management this can be advantageous if, instead of forecasting the entire density, interest is restricted to any kind of risk measure that only takes the downside (or upside) of the predicted distribution into account, such as the VaR, or expected shortfall.[1]

One source of inaccuracies in the prediction of downside risk in this context is likely to stem from asymmetries in the data which are not adequately captured by the chosen model, even when the model allows for (i) an asymmetric response to shocks in the GARCH equation and (ii) a flexible asymmetric innovations distributional assumption. This is the case, for example, with the asymmetric-power-GARCH (APARCH) model coupled with a generalized asymmetric Student's t distribution, denoted GAt-APARCH, which has been shown independently by Mittnik and Paolella (2000) and Giot and Laurent (2004) to deliver relatively (compared to other models) very accurate VaR forecasts. And in fact, several extensive empirical studies by Paolella and Steude (2007) show that with weighting functions designed to place more weight on negative observations, this asymmetry problem can be mitigated and highly accurate VaR forecasts can be obtained. We refer to this method as tail weighted ML (or, in short, TaWML).

In the current context, we consider time series models for financial asset returns and propose a third way to generalize the usual likelihood paradigm of implicitly placing equal weight on each of the observations in the data sample. The idea behind the weighing method we propose in this paper is the following: Assuming that financial time series consist of a set of data points that do not yield any advantage for predicting the future path of the true DGP – possible stemming from noise traders or other forms of destructive outlier formation –, then the standard maximum likelihood principle is essentially overestimating the usefulness of these data points by allocating the same weight across all data points equally. From an economic perspective, for example high volume days, in which many market participants agree on one price, might convey more information about the true DGP (possible also because the disruptive impact of noise traders goes by unnoticed and without any major effect on the price level). In this case standard ML *under*-estimates the importance of those values, again, because it assumes implicitly placing equal weights for all datapoints.

The aim of this paper is to propose a new method of retrieving more relevant information from the data by differentiating the weights within the sample. The proposed method, volume weighted maximum likelihood, or short VoWML, is based on external variables, for example, trading volume, trading liquidity or news arrivals to detect data points that convey more information about the time series' future (or generally its DGP) than the average data point.

It is also appealing that the proposed method neatly nests all known weighting schemes, but more importantly, while TiWML and TaWML both address different shortcomings of the time series models at hand, our proposed method should improve the forecasting ability even if one knew the true DGP. Hence, the three weighting schemes, TiWML, TaWML and VoWML can also be combined, also given that they address different shortcomings of the equally weighted case and we show below that this indeed yields further improvement compared to the use of only one of the weighting schemes.

There are now numerous ways of computing the VaR of a particular financial asset, with comparisons and extensions of some of the most promising methods detailed in Bao et al. (2006) and Kuester et al. (2006). As emphasized in the methodology developed in Hartz et al.

---

[1] In spite of the (well) known criticism of the VaR, we report the VaR in this paper because of its high relevance in practice and because other risk measures such as the expected shortfall are calculated using the VaR as a necessary input.

Model	VaR–level	DOW		Dax		FTSE		Nikkei		NASDAQ	
		$\delta = 1$	min	$\delta = 1$	min	$\delta = 1$	min	$\delta = 1$	min	$\delta = 1$	min
normal-GARCH	1%	0.508	0.508	0.503	0.502	0.503	0.503	0.534	0.533	0.506	0.501
	2.5%	1.293	1.278	1.279	1.066	1.279	1.271	1.253	1.167	1.248	1.175
	5%	2.516	2.359	2.480	1.250	2.480	2.376	2.362	1.928	2.240	1.874
GAt-GARCH	1%	0.452	0.432	0.397	0.390	0.315	0.277	0.534	0.466	0.465	0.451
	2.5%	1.165	1.128	0.921	0.899	0.881	0.760	1.259	0.984	1.100	1.046
	5%	2.250	2.190	1.690	1.165	1.803	1.572	2.006	1.512	1.960	1.795

Table 1: Mean squared deviation (MSD) of the empirical from the theoretical tail probability for WS-4. The entry "$\delta = 0$" contains the MSD for the equal weighted case (as a reference) and the entry named "min" stands for the minimal achieved MSD when more weight is given to the observations. For all data sets δ_i are between 0 and 15.

(2006), the search for ever-more-complicated models might not be as fruitful as starting with simpler, easily-estimated models and changing instead something other than the parametric form of the model. In particular, those authors demonstrate the viability of using the bootstrap and a bias-correction step in conjunction with the rudimentary (and otherwise inadequate, but easily estimated) normal-GARCH model. In this paper, we use several easily estimated parametric GARCH models, including the normal-GARCH and GAt-APARCH, in conjunction with the weighted likelihood schemes.[2] We expect, and confirm, that the weighting schemes will improve all the models and improve the simplest model (normal-GARCH) the most, but the best results are obtained using the more sophisticated models.

In this paper, we restrict attention to a single asset. For use with a portfolio, the usual method of constructing a univariate series of returns can be used, based on the returns of the components of the portfolio; see, e.g., Dowd (2005, Section 4.1) for discussion. An extension of the TiWML weighting methodology to the multivariate case is immediate, while for TaWML and VoWML, several ideas could be entertained and will be pursued in future work.

The remainder of this paper is organized as follows. Section 2 briefly summarizes the two weighted likelihood families, TiWML and TaWML and introduces the new method VoWML. A short description of the parametric models that are used for illustrative purpose is contained in Section 3, while Section 4 lists the model selection criteria used within empirical Section 5. Section 6 contains the conclusion and outlook.

2 Weighting Families

There are currently two distinct ways of using weighted ML known to the literature. The first one, we will refer to as TiWML. Its fundamental concept is to place more weight on recent observations. Let T be the length of the time series under study. To implement the weighting scheme, a vector of weights $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_T)$ is used such that it is standardized to sum to a constant, such as T (as with the conventional MLE) or to one, $\sum_{t=1}^{T} \tau_t = 1$. The model parameters are then estimated by maximizing the weighted likelihood, whereby the likelihood component associated with period t is multiplied by τ_t. In Mittnik and Paolella (2000), the two weighting schemes, geometric and hyperbolic, are proposed and studied, given respectively by $\tau_t \propto \rho^{T-t}$ and $\tau_t \propto (T - t + 1)^{\rho-1}$, where the single parameter ρ dictates the shape of the weighting function. In both cases, values of $\rho < 1$ ($\rho > 1$) cause more recent observations to be given relatively more (less) weight than those values further in the past; $\rho = 1$ corresponds to standard ML estimation.

The second way of leaving the traditional assumption of qual weighting is by placing relatively more weight on negative (or, possibly, only on the *extreme* negative) returns in the sample, introduced by Paolella and Steude (2007). When interest centers on the downside risk potential of a financial position and whole density forecasts are not necessarily needed, TaWML can lead to considerable forecasting improvement. As in the time weighting scheme, each of the T components of the likelihood gets multiplied by the tth component of a weight

---

[2] A completely different approach to modelling in this context is to use nonparametric methods. Kuester et al. (2006) demonstrate that, even for the successful nonparametric methods (which, interestingly, all have, as part of their methodology, a parametric component, including EVT-GARCH and filtered historical simulation), the choice of the parametric modelling component associated with the GARCH filter and distributional assumption for the innovations is decisive for the out–of–sample performance. As such, parametric methods still appear to be of great relevance in this context and we restrict attention to them in this paper.

vector $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_T)$, with $\sum_{t=1}^{T} \omega_t = 1$. Paolella and Steude (2007) also test for a variety of strategies placing more weight on negative observations. These weighting schemes differ mostly in the number of location and shape parameters. In the most effective ones are shown in Figure 1. The first scheme is structured as follows: split the returns into two groups of negative and positive observations, and assign an equal, but relatively higher, weight to the negatives, and an equal (but smaller) weight to the positives. This can be graphically seen in the right panel of Figure 1, labelled "Weighting Scheme 1", subsequently referred to as WS-1. Let w_n (w_p) be the weight on the negative (positive) values, of which there are n_n (n_p) of them. In order for the weights to sum to one, we require $w_n n_n + w_p n_p = 1$. In the standard (unweighted) setup, clearly $w_n = w_p = 1/(n_n + n_p)$, and to characterize the weighted case, we use parameter δ and take

$$w_n = \frac{\delta}{n_n + n_p}, \quad w_p = \frac{1 - w_n n_n}{n_p}, \quad \text{where} \quad \delta \in \left[0, \frac{n_n + n_p}{n_n}\right].$$

Observe that, if $\delta = 0$, then $w_n = 0$ and $w_p = n_p^{-1}$, i.e., all weight is placed on positive observations, and if $\delta = (n_n + n_p)/n_n$, then $w_n = n_n^{-1}$, $w_p = 0$, and all weight is placed on the negative observations. For financial returns data, $n_n \approx n_p$, so the upper bound on δ is approximately two. Given δ, values w_n and w_p are determined, from which $\boldsymbol{\omega}$ can be constructed. The optimal value of δ for a particular data set needs to be determined empirically, and its computation is discussed below.

A second important weighting scheme (middle right subplot of Figure 1), denoted WS-3, makes use of a continuous, strictly decreasing function on $[0,1]$ such that, the more negative the innovation, the more weight it receives. In particular, WS-3 is based on a cumulative distribution function (cdf) of a random variable with finite support suggests itself. Because of its flexibility and availability in virtually all statistical computing platforms, the use is the Beta(p,q) cdf, given by the incomplete beta ratio $F_{\text{Beta}}(x; p, q) = \bar{B}_x(p, q)$, and where the choices of p and q are discussed below. For WS-3, first construct $\pi_t = 1 + r_t / \max_t(-r_t)$, and, for $\delta \geq 0$, let

$$\omega_t = \begin{cases} \delta(1 - F_{\text{Beta}}(\pi_t; p, q)) + 1, & \text{if } r_t < 0, \\ 1, & \text{otherwise,} \end{cases} \quad (1)$$

where $\delta \geq 0$ determines how much weight is assigned to negative returns, with a value of zero yielding the default case of equal weights. In the final step the weights are standardized to sum to one. This results in a continuous weighting scheme, as illustrated in Figure 1.[3]

The third and new method allocates the weights according to the realizations of external variables that yield information about the current state of the DGP. Similar to TaWML one can now assume a variety of weighting schemes. The basic building block can be a pdf or cdf, depending on the actual time series under study and most importantly, the external variable that serves as the weight determining factor. Similar to TaWML and starting with a daily setting, the model parameters are estimated by maximizing the weighted likelihood, whereby the likelihood component associated with period t is weighted according to the external variable in period t, not like in TaWML according to the r_t of that period. For example if one uses the trading volume on day t, v_t, Equation 1 for WS-3 will only change with respect to r_t vs. v_t, all else staying the same. So all weighing schemes of TaWML can also be applied to VoWML, only substituting r_t with the external variables at hand.

---

[3]WS-4 is constructed according to Equation 1 using a PDF instead of a CDF.

In principle, one could treat the three variables δ, p and q as tuning parameters to be empirically determined, though this would involve a prohibitively large amount of computation. Instead, via trial and error over several data sets, we suggest using the fixed values $p = 250$ and $q = 2$ and optimizing only over δ (as discussed below). Changing the values of p and q in a moderate manner and recomputing the optimal value of δ resulted in very similar forecasting quality.

Note that, for all three TiWML, TaWML and VoWML, the optimal value of the tuning parameter (either ρ, δ, v) cannot be estimated with the model parameters by maximizing the likelihood, but must be obtained with respect to some criterion outside of the likelihood function of the T observations. We make use of out–of–sample VaR predictions for this purpose, as discussed in more detail below.

3 Parametric Models for Computing Value at Risk

To demonstrate the value of the aforementioned weighting schemes, we explore how the forecasting ability of different parametric models is influenced by their use. In this first empirical study on the new method we entertain three different models which increase in their degree of sophistication and overall ability for downside risk prediction.

1. Standard GARCH(1,1) with normally distributed innovations (normal–GARCH) and an AR(1) term. That is, the asset or portfolio return at time t, r_t, is assumed to follow the process

$$r_t = a_0 + a_1 r_{t-1} + \epsilon_t, \tag{2}$$

   where $\epsilon_t = \sigma_t Z_t$, the Z_t are i.i.d. standard normal innovations and

$$\sigma_t^2 = c_0 + c_1 \epsilon_{t-1}^2 + d_1 \sigma_{t-1}^2. \tag{3}$$

2. GARCH(1,1) with Student's t innovations (t–GARCH). This is the same model as in (2) and (3), but with the Z_t being i.i.d. Student's t.

3. GARCH(1,1) with innovations from the generalized asymmetric Student's t distribution (GAt-GARCH). The model is again given by (2) and (3), but with Z_t i.i.d. GAt, with density $f(z; d, \nu, \theta)$ given by

$$C \left(1 + \frac{(-z\theta)^d}{\nu} \right)^{-\frac{\nu+1}{d}} \mathcal{I}(z < 0) + C \left(1 + \frac{(z/\theta)^d}{\nu} \right)^{-\frac{\nu+1}{d}} \mathcal{I}(z \geq 0), \tag{4}$$

   where $d, v, \theta > 0$, $\mathcal{I}$ is the indicator function and the constant of integration is given by $C = [(\theta + \theta^{-1}) d^{-1} \nu^{1/d} B(d^{-1}, \nu)]^{-1}$, where B is the beta function.[4]

4 Model Comparison Criteria

With risk management in mind we analyze the forecasted VaR performance as well as a number of other statistics for evaluating the quality of the predicted density.

---

[4] Expressions for the moments, cdf, and expected shortfall of the GAt are straightforwardly derived; see, e.g., Paolella (2007, Exercise 7.7).

		All News Counts		News Counts, REL=.9	
Model	VaR–level	$\delta = 1$	min	$\delta = 1$	min
normal-GARCH	1%	0.538	0.538	0.538	0.538
	2.5%	1.291	1.290	1.291	1.288
	5%	2.521	2.481	1.521	1.455
GAt-GARCH	1%	0.488	0.444	0.488	0.429
	2.5%	1.179	1.549	1.179	1.136
	5%	2.295	1.902	2.295	1.859

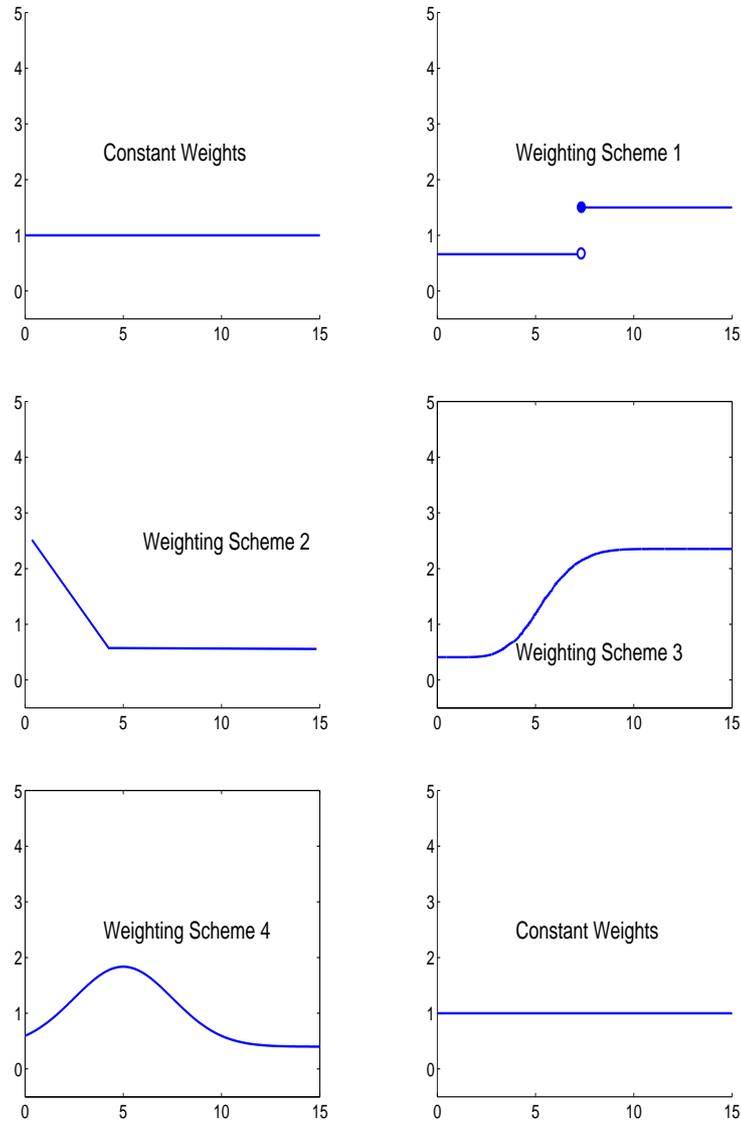Table 2: Same as in Figure 1 but for different methods of news counts.

Figure 1: Different possible weighting schemes through which more weight is placed on negative observations in the sample. On the horizontal axes are the returns, the vertical axes shows the weights associated with them. Weighting scheme "Constant Weights" corresponds to the default case of equal weights on negative and positive returns.

First, we compare the Anderson-Darling (AD) and the Cramér-von Mises (CM) test statistics based on N consecutive out-of-sample forecasts to assess density forecast quality.

Let $\hat{\boldsymbol{p}}$ be the series of realized predictive cdf values evaluated at the true, observed returns, $\hat{p}_i = \hat{F}_{i|\mathcal{F}_{i-1}}(\varepsilon_i; \hat{\boldsymbol{\theta}}_{i-1})$, $i = 1, \ldots, N$, and let $\hat{\boldsymbol{p}}^{[s]}$ denote the sorted vector, $\hat{p}_1^{[s]} \leq \hat{p}_2^{[s]} \leq \cdots \leq \hat{p}_N^{[s]}$. Both test statistics measure the deviation from uniformity, given by

$$\text{AD} = -N - \sum_{i=1}^{N} \frac{2i-1}{N} \left(\log(\hat{p}_i^{[s]}) + \log(1 - \hat{p}_{N-i+1}^{[s]}) \right)$$

and

$$\text{CM} = \frac{1}{12N} + \sum_{i=1}^{N} \left(\frac{2i-1}{2N} - \hat{p}_i^{[s]} \right)^2,$$

respectively.

Second, as a proxy for the iid property we check for serial correlation in $\hat{\boldsymbol{p}}$ using the Ljung-Box (LB) statistic,

$$\text{LB} = N(N+2) \sum_{i=1}^{m} \frac{\hat{\rho}_i^2}{N-i},$$

where $\hat{\rho}_i$ is ith autocorrelation from the m-lagged sample autocorrelation function, and $m = 20$.

Third, and as it is (still) more common in risk management literature we also study the empirical violation probabilities associated with the VaR forecasts (which, arguably, is the measure which first needs to be accurately fulfilled before others should even be considered; see Christoffersen and Pelletier, 2004; Kuester et al., 2006; and the references therein).

Comparing the VaR forecasting quality, we look at the predicted VaR as well as approaches by Kuester 2006 and Christoffersen 1998. Similar to Kuester 2006 and , we use a simple measure based on the coverage results for all VaR levels up to $100\lambda\%$. The basic idea behind this statistic is to compare the deviation of the forecast cdf from the uniform cdf and, as such, it captures the excess of percentage violations over the actual VaR level. The deviation is defined as $100(F_U - \hat{F}_e)$, where F_U is the cdf of the standard uniform random variable. $\hat{F}_e$ refers to the empirical cdf formed from $\hat{\boldsymbol{p}}$. An interesting and quite convenient way of comparing forecasting qualities among different set ups, models and data sets is reported in Paolella (2008) and we also report their integrated root mean squared error (IRMSE), summed up over the left tail up to the VaR level of interest. In fact, the IRMSE is closely related to the CM statistic but with the sum truncated at $h = \lceil \lambda N \rceil$, i.e.,

$$\text{IRMSE} = \sqrt{\frac{1}{h} \sum_{i=1}^{h} \left(100\frac{2i-1}{2N} - 100\hat{p}_i^{[s]} \right)^2}.$$

And finally similar to report the hit sequence of realized predictive VaR violations,

$$v_i = \mathbf{1}_{\varepsilon_i \leq q}, \quad q = \text{VaR}(\lambda, i), \tag{5}$$

whith $\mathbf{1}$ being the indicator function. Given the true model, the v_i are iid Bernoulli(λ). Based on this sequence, the usual likelihood-ratio (LR) test statistic is computed as proposed in Christoffersen 1998,

$$\text{LR}_{\text{CC}} = \text{LR}_{\text{UC}} + \text{LR}_{\text{IND}},$$

where LR_{UC} and LR_{IND} refers to the unconditional coverage and the independency property, respectively.

5 Data and Empirical Results

We present our results for six time series of daily financial returns data: Dow Jones 30, Dax 30, FTSE 100, Nikkei 225 and the NASDAQ Composite Index, which comprises about 2000 stocks varying slightly with the reporting date. Prices and volumes are taken from Thomson Reuters Datastream. The stock indices range from January 1st, 2003 to January 1st, 2011, matching the time span of the Thomson Reuters NewsAnalytics Database from which the company news counts and news sentiment are retrieved. We calculate continuously compounded percentage returns, $r_t = 100 \left(\log P_t - \log P_{t-1}\right)$, where P_t denotes the index level at time t. For volume and news frequency data we calculate first differences.

The data about news releases including news sentiment information are taken from the innovative Thomson Reuters NewsAnalytics database. Starting in January 2003 this database contains all news wire stories from the Thomson Reuters News service which sums up to several thousand news a day. More interesting though for our purpose are the pre-processed indicator variables of the News Analytics database that are attached by Thomson Reuters to each news story. Among them are

- a broad measure of the topic of the news item,

- the novelty of the news item,

- the relevance of the news item for the company, and,

- the sentiment or tone of the news item, e.g., News Analytics rates each news according to positive, neutral and negative author sentiment, and

- a proxy for the likelihood that the given sentiment score is correctly specified.

Clearly, besides the number of news each day, for our purpose it is the sentiment of the news items that is most useful. NewsAnalytics electronically rates each news as either positive, neutral or negative, or 1, 0 and -1, respectively. Also of great interest is the indicator variable *relevance* which measures how targeted the story might be to a company mentioned in the news. Values range in $[0, 1]$ and a value of one is assigned for example when the company is mentioned in the headline.

News Analytics attaches these indicator variables to each news item based on an electronic algorithm that makes use of linguistic pattern recognition analysis.

5.1 Empirical Results for Trading Volume

We will first look at the results on trading volume followed by the analysis of news arrivals and investor sentiment. For each series and model, we use a moving window of length $1,000$ days with parameter re–estimation for each window. (Arguably, we could have used growing windows of data, but a moving window of fixed length was used to allow more consistent comparisons across models and data sets.) This results in roughly 8,000 out of sample forecasts for each model combination.

Turning to trading volume, consider the weighting scheme, WS-4. After some initial experimentation, we chose a grid of δ–values of $[0, 15]$ and used grid steps of 0.1. Values for $\delta > 0$ mean more weight is placed on daily observations that have a "normal" trading volume. One can call them the average trading volume days (see again Figure 1); $\delta = 0$ is the standard

non–weighted case. Shape parameters for WS-4 are set to $p = 15$ and $q = 25$ but the results are qualitatively the same for slight variations of those for any dataset and model.

As described above, Figure 2 contains a graphical depiction of the accuracy of the one-step VaR predictions, as a function of the weighting parameter δ, using the MSD measure for the VaR level of up to 5%. The figure shows the results for the GAt-APARCH. The first important result for the Dow Jones data set is that, the forecasting performance is improved by allocating more weight to average trading volume observations. This shows that even the most sophisticated model we entertain (the asymmetric GAt-GARCH model) can be improved upon by placing more weight on the medium observations. Unsurprisingly, the behavior of the MSD differs among the models. Notably, within the range of δ we used, the MSD for some models worsens after some optimal δ–value is reached, while for other models, the turning point seems to lie outside the applied range. For example, in Figure 2 it can be seen that increasing δ up to a value of about 10, the forecasting by our proposed method improves steadily and then after the optimal point it flips back.

To save space and further condense the results, Table 2 gives an overview of the MSD of the other data sets for the normal and GAt-GARCH models. (The graphs for the other data sets are qualitatively very similar to those in Figures 2.) For a given data set and ξ_{max} level, the table contains the MSD value for $\delta = 0$ and the lowest MSD that was obtained in the δ–interval $[0, 15]$. For example, for the normal-GARCH model and $\xi_{max} = 5\%$, the MAD for the Dow Jones index in the equal–weight case is of 2.52, and 2.36 in the optimal weighted case. For all data sets, we see that the most improvement occurs for the simplest model (normal-GARCH), while the best results are obtained for the more elaborate GARCH models, in which case there is still improvement from use of the weighting scheme, though less than with the normal-GARCH model. Another (perhaps) surprising result is that our weighting scheme across models and datasets yields more improvement for the 5% VaR than for the 1% VaR.

Turning to news arrivals, Table 1 gives an overview of the out of sample forecasting performance when applying WS-4 for news arrivals. As the table shows the results for the Dow Jones, it is the number of news about Dow Jones companies that we filter out from the Thomson Reuters NewsAnalytics database. First you see that the better the model, the better the forecast (as in the volume weighted case), but you see a little less room for model improvement when using news counts vs. trading volume. However, as described above in the Thomson Reuters database there is a qualifier called "relevance" that indicates how relevant a certain message is for the company at hand. The higher now the relevance score of the news (instead of just taking all US news that get distributed over the Thomson Reuters news channel for the Dow Jones companies) the bigger the forecasting improvement.

It is interesting to see that all the above results concerning the appropriateness of the proposed method are also valid when looking at other measures of forecasting quality. Results are qualitatively the same for the Anderson-Darling and the Cramér-von Mises test statistics.

6   Conclusions and Outlook

Risk Management involves optimizers among which the Maximum Likelihood Estimator (MLE) is the most popular. Because different data observations in the sample yield different information about the true data generating process, individual weights should also be allocated.

However, the traditional MLE ignores this and assume equal likelihood weights for all data points. In this paper we propose several potential weighting schemes to do this, confirming that large gains in accuracy can be realized, even when using sophisticated GARCH models which account for asymmetry and or the leverage effect.

There are several ways of promising future work. One involves better proxies of investor sentiment trying to harvest behavioral biases in the data. The Thomson Reuters NewsAnalytics database also contains the sentiment of each news and it is ongoing work to test the influence on this variable. Also, of great interest to practical risk management is to extend our setting to allow for a multivariate and (possibly) high frequency world.

The determination of the optimal values of the tuning parameters which dictate the weighting functions, δ and ρ, clearly entail a significant amount of computation and would prevent the full method from being used on a daily basis. However, the values of the weight parameters need not be re-estimated every day, and, once found, should prove to be superior to using the default values corresponding to the equally-weighted, standard likelihood setting. Once values for δ and ρ have been determined, the estimation and VaR prediction of the model takes no longer to compute than using the standard likelihood.

References

Amisano, G. and Giacomini, R. (2006). Comparing density forecasts via weighted likelihood ratio tests. *Forthcoming: Journal of Business and Economic Statistics*.

Boudoukh, J., Richardson, M., and Whitelaw, R. F. (1998). The best of both worlds: A hybrid approach to calculating value at risk. *Risk*, 11(5):64–67.

Cheng, T.-C. (2005). Robust regression diagnostics with data transformations. *Computational Statistics and Data Analysis*, 49(3):875–891.

Christoffersen, P. F. and Pelletier, D. (2004). Backtesting value–at–risk: A duration–based approach. *Journal of Financial Econometrics*, 2:84 – 108.

Diebold, F. X., Gunther, T. A., and Tay, A. S. (1998). Evaluating density forecasts with applications to financial risk management. *International Economic Review*, 39(4):863–883.

Dowd, K. (2005). *Measuring Market Risk*. Wiley, New York, 2nd edition.

Giot, P. and Laurent, S. (2004). Modelling daily value–at–risk using realized volatility and arch type models. *Journal of Empirical Finance*, 11:379–398.

Hadi, A. S. and Luceno, A. (1997). Maximum trimmed likelihood estimators: a unified approach, examples, and algorithms. *Computational Statistics and Data Analysis*, 25:251–272.

Hartz, C., Mittnik, S., and Paolella, M. S. (2006). Accurate value–at–risk forecasting based on the normal–garch model. *Computational Statistics & Data Analysis*, 51:2295–2312.

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

Markatou, M. (2000). Mixture models, robustness, and the weighted likelihood methodology. *Biometrics*, 56:483–486.

Mittnik, S. and Paoella, M. S. (2000). Conditional density and value-at-risk prediction of asian currency exchange rates. *Journal of Forecasting*, (19):313–333.

Mittnik, S. and Paolella, M. S. (2000). Conditional density and value-at-risk prediction of asian currency exchange rates. *Journal of Forecasting*, 19:313–333.

Paolella, M. S. (2007). *Intermediate Probability: A Computational Approach*. John Wiley & Sons, Chichester.

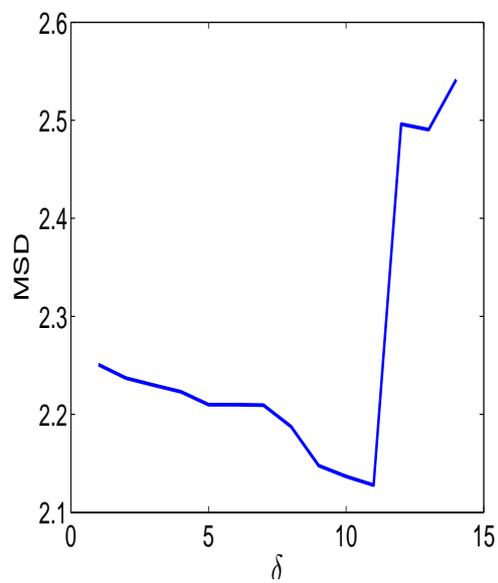Tay, A. S. and Wallis, K. F. (2000). Density forecasting: A survey. *Journal of Forecasting*, 19:235–254.

Figure 2: Mean squared deviation (MSD) of the empirical from the theoretical tail probability for WS-4 for the Dow Jones index. The entry "$\delta = 0$" contains the MSD for the equal weighted case (as a reference).