Submitted to the Annals of Applied Statistics

THE ROLE OF THE INFORMATION SET FOR FORECASTING – WITH APPLICATIONS TO RISK MANAGEMENT

By Hajo Holzmann* and Matthias Eulert

Department of Mathematics and Computer Science, Philipps-University Marburg, Germany

Predictions are issued on the basis of certain information. If the forecasting mechanisms are correctly specified, a larger amount of available information should lead to better forecasts. For point forecasts, we show how the effect of increasing the information set can be quantified by using strictly consistent scoring functions, where it results in smaller average scores. Further, we show that the classical Diebold–Mariano test, based on strictly consistent scoring functions and asymptotically ideal forecasts, is a consistent test for the effect of an increase in a sequence of information sets on h-step point forecasts. For the value at risk (VaR), we show that the average score, which corresponds to the average quantile risk, directly relates to the expected shortfall. Thus, increasing the information set will result in VaR forecasts which lead on average to smaller expected shortfalls. We illustrate our results in simulations and applications to stock returns for unconditional versus conditional risk management as well as univariate modeling of portfolio returns versus multivariate modeling of individual risk factors. The role of the information set for evaluating probabilistic forecasts by using strictly proper scoring rules is also discussed.

1. Introduction. Making and evaluating statistical forecasts is a basic task for statisticians and econometricians. While probabilistic forecasts, consisting of a complete predictive distribution, are most informative (cf. Gneiting, Balabdaoui and Raftery, 2007), interest often focuses on single-value point forecasts (Gneiting, 2011). For example, in quantitative risk management, the goal is to estimate certain functionals of a predictive distribution such as the value at risk (VaR) or the expected shortfall (McNeil, Frey and Embrechts, 2005).

Forecasts are issued on the basis of certain information. Evidently, increasing the information set should lead to better forecasts, at least if the forecasting mechanisms are correctly specified. We shall call such forecasts ideal. In this article, we show how an improvement of ideal forecasts by increasing the information set can be quantified by using strictly consistent scoring functions (Gneiting, 2011), where it results in smaller average scores. Further, we show that the classical Diebold and Mariano (1995) test, based on strictly consistent scoring functions and asymptotically ideal forecasts, is a consistent test for the effect of an increase in a sequence of information sets on h-step point forecasts.

As a most important example, consider evaluating VaR forecasts. Formally, the VaR is a (high, say 0.99 or 0.999) quantile of the loss distribution. Unconditional methods base the VaR on the unconditional distribution of the risk factors thus using a trivial information set, while conditional methods refer to a conditional distribution typically given the historical data, see McNeil, Frey and Embrechts (2005). For conditional methods, the information set

*Address for correspondence: Prof. Dr. Hajo Holzmann, Philipps-Universität Marburg, Fachbereich Mathematik und Informatik, Hans-Meerwein-Straße. 35032 Marburg, Germany, Fon: +49 6421 2825454

Keywords and phrases: forecast, information set, scoring function, scoring rule, value at risk

may vary as well: in a portfolio point of view it only includes the portfolio returns, while a modeling of the individual risk factors involves a larger information set.

Unconditional backtesting consists in checking whether the relative frequency of exceedances of the VaR estimates corresponds to the level of the VaR. This is, as the name suggests, satisfied by both unconditional and conditional methods if correctly specified. Conditional methods are accompanied in case of one-step ahead estimates by checking whether exceedances of VaR forecasts occur independently (the i.i.d. hypothesis, c.f. Christoffersen, 1998; McNeil, Frey and Embrechts, 2005). However, independence of exceedance indicators alone does not adequately take into account the size of the information set for the conditional methods, see also Berkowitz, Christoffersen and Pelletier (2011).

We show that by evaluating (ideal) VaR forecasts by scoring functions, one can distinguish between VaR forecasts arising from distinct information sets. Interestingly, increasing the information set will result in VaR forecasts which lead to smaller expected shortfalls, unless an increase in the information set does not result in any change in the VaR forecast.

The paper is organized as follows. The general methodology is developed in Section 2. To illustrate, we start in Section 2.1 with an example from regression analysis. We recall the well-known fact that by including additional variables and thus increasing the information set, the mean-squared prediction error of the (population, i.e. ideal) mean regression function is reduced. Then, turning to general expectile regression, we indicate that our subsequent results imply that including additional variables will reduce the mean asymmetric squared loss of the (ideal) expectile-regression functions. In Section 2.2 we show how the effect of a larger information set for issuing a certain point forecast can be quantified by using strictly consistent scoring functions. Section 2.3 is concerned with the same problem in case of evaluating probabilistic forecasts by using proper scoring rules. See also the note by Tsyplakov (2011), which comments on the paper by Mitchell and Wallis (2011) which in turn is a critical comment on Gneiting, Balabdaoui and Raftery (2007). In Section 2.4, we investigate the properties of the Diebold and Mariano (1995) test in the situation of nested sequences of information sets and asymptotically ideal forecasts.

Section 3 contains a detailed discussion of methods to evaluate VaR forecasts. We start by discussing applications of the VaR such as risk controls for trading desks, VaR-based portfolio choice and regulatory uses, as well as general strategies for issuing VaR forecasts. In Section 3.1, we focus on exceedance indicators which are the typical tool for backtesting VaR forecasts, and in Section 3.2 we turn to the quantile loss (the strictly consistent scoring function for the VaR), and relate its expected value to the expected shortfall.

In Section 4, we conduct a simulation study and give applications to series of stock-returns for value at risk estimation, when comparing first unconditional versus conditional methods and second univariate modeling on the basis of portfolio returns versus multivariate modeling of the individual risk factors. Section 5 concludes, while technical proofs are deferred to an appendix.

2. Quantifying the role of the information set.

2.1. *An introductory example from regression analysis.* To motivate the upcoming discussion, consider an example in a regression framework. Suppose that a triple (Y, X_1, X_2) of random variables is observed, where Y is the dependent variable with $E|Y| < \infty$, and X_1, X_2 are explanatory random variables.

Consider the mean regression $g(x_1, x_2) = E(Y|X_1 = x_1, X_2 = x_2)$ of Y on (X_1, X_2), as well as $f(x_1) = E(Y|X_1 = x_1)$ of Y on X_1 only. Given values x_1, x_2, in which sense is $g(x_1, x_2)$ a more precise forecast than $f(x_1)$ for the conditional mean of Y, or phrased otherwise, in which sense is the forecast improved if the information set is increased from $\mathcal{F} = \sigma(X_1)$ to $\mathcal{G} = \sigma(X_1, X_2)$?

As is well-known, if $EY^2 < \infty$ we have that P-almost surely (P–a.s.),

$$\begin{aligned} E\big((Y - g(X_1, X_2))^2|X_1\big) &= E\big(Y^2|\mathcal{F}\big) - E\big((E(Y|\mathcal{G}))^2|\mathcal{F}\big) \\ &\le E\big(Y^2|\mathcal{F}\big) - (E(Y|\mathcal{F}))^2 = E\big((Y - f(X_1))^2|X_1\big) \end{aligned}$$

since by the conditional Jensen inequality, $(E(Y|\mathcal{F}))^2 \le E\big((E(Y|\mathcal{G}))^2|\mathcal{F}\big)$, and therefore also that the unconditional squared forecast error is reduced

$$E\big((Y - g(X_1, X_2))^2\big) \le E\big((Y - f(X_1))^2\big).$$

Patton and Timmermann (2012) discuss the special case of mean prediction and the effect of an increased information set in a dynamic context.

Now, the natural question is whether analogous statements are true if we move away from the simple mean regression, say to an expectile regression on the α expectile, $\alpha \ne 1/2$, or even consider the whole predictive distributions $\mathcal{L}(Y|\mathcal{F})$ and $\mathcal{L}(Y|\mathcal{G})$.

Recall that the α expectile τ_α of a distribution function F on $\mathbb{R}$ with finite first moment is defined as the unique solution in τ to

$$\alpha \int_\tau^\infty (y - \tau)\, dF(y) = (1 - \alpha) \int_{-\infty}^\tau (\tau - y)\, dF(y).$$

Let $g_\alpha(x_1, x_2)$ (resp. $f_\alpha(x_1)$) denote the α expectile of the conditional distribution function of Y given $X_1 = x_1, X_2 = x_2$ (resp. given $X_1 = x_1$). Our result below implies that if $EY^2 < \infty$, and if we replace the squared loss $(y - m)^2$ for the mean by the asymmetric squared loss $S_\alpha(y, \tau) = |1_{\tau \ge y} - \alpha|\, (y - \tau)^2$ for the α expectile, then P–a.s.

$$E\big(S_\alpha(Y, g_\alpha(X_1, X_2))|\mathcal{F}\big) \le E\big(S_\alpha(Y, f_\alpha(X_1))|\mathcal{F}\big),$$

as well as

$$E\big(S_\alpha(Y, g_\alpha(X_1, X_2))\big) \le E\big(S_\alpha(Y, f_\alpha(X_1))\big)$$

with equality if and only if $g_\alpha(X_1, X_2) = f_\alpha(X_1)$. This will be deduced by using the fact that the above loss functions are strictly consistent for the functionals, as defined below.

2.2. *Functionals and scoring functions.* We start by recalling the concept of strictly consistent scoring functions, see Gneiting (2011). Let Θ be a class of distribution functions on a closed subset $D \subset \mathbb{R}$, which we identify with their associated probability distributions, and let $T : \Theta \to \mathbb{R}$ be a (one-dimensional) statistical functional. We let $\mathcal{B}(\Theta)$ denote the Borel σ-algebra on Θ w.r.t. the topology of weak convergence of distribution functions (or probability measures), and we let $\mathcal{B}$ denote the ordinary Borel σ-algebra on $\mathbb{R}$. We shall call the functional T measurable if it is $\mathcal{B}(\Theta) - \mathcal{B}$-measurable.

A *scoring function* is a measurable map $S : \mathbb{R} \times D \to [0, \infty)$. Then $S(x, y)$ is interpreted as the loss if forecast x is issued and y materializes. S is consistent for the functional T relative to the class Θ, if

$$\text{for all } x \in \mathbb{R},\ F \in \Theta : \quad E_F\big(S(T(F), Y)\big) \le E_F\big(S(x, Y)\big),$$

where Y is a random variable with distribution function F, and we assume that the relevant expected values exist and are finite. Thus, the true functional $T(F)$ minimizes the expected loss under F. If

$$E_F\big(S(T(F),Y)\big) = E_F\big(S(x,Y)\big) \quad \text{implies that} \quad x = T(F),$$

then S is *strictly consistent* for T. If the functional T admits a strictly consistent scoring function, then it is called *elicitable* (relative to the class Θ). For several functionals such as mean, quantiles and expectiles Gneiting (2011) characterizes all strictly consistent scoring functions which additionally satisfy

$$\begin{aligned} &1. \quad S(x,y) \geq 0 \text{ with equality if and only if } x = y,\\ &2. \quad S(x,y) \text{ is continuous in } x \text{ for all } y \in D,\\ &3. \quad \text{the partial derivative } \partial_x S(x,y) \text{ exists and is continuous in } x \text{ for } x \neq y. \end{aligned} \tag{1}$$

Note that for simplicity we do not consider set-valued functionals. Our results could be extended to include these, but the formulations would become more cumbersome. Thus, in case of quantiles, we assume that all distributions functions in Θ are strictly increasing.
Gneiting (2011) also points out that well-known functionals such as variance or expected shortfall are not elicitable. Heinrich (2013) obtains a corresponding negative result for the mode functional, despite the convexity of the level sets for the mode.
Now let us consider a forecasting situation. Forecasts are issued on the basis of certain information. Let $(\Omega, \mathcal{A}, P)$ be a probability space, let $\mathcal{F} \subset \mathcal{A}$ be a sub-σ-algebra of $\mathcal{A}$ (the information set), and let $Y : \Omega \to \mathbb{R}$ be a random variable. The aim is to predict a particular functional of the conditional distribution of Y given $\mathcal{F}$.

THEOREM 1. *Let $F_{Y|\mathcal{F}}(\omega,\cdot)$ be the conditional distribution function of Y given $\mathcal{F}$. Assume that for each $\omega \in \Omega$, $F_{Y|\mathcal{F}}(\omega,\cdot) \in \Theta$. If $T : \Theta \to \mathbb{R}$ is measurable, then $T(F) = T\big(F_{Y|\mathcal{F}}(\omega,\cdot)\big) = \hat{Y}(\omega)$ is an $\mathcal{F}$-measurable r.v. If T is elicitable (over Θ) and if S is a strictly consistent scoring function for T, then for any $\mathcal{F}$-measurable r.v. Z, we get*

$$E\left(S(\hat{Y},Y)|\mathcal{F}\right)(\omega) \leq E\left(S(Z,Y)|\mathcal{F}\right)(\omega) \qquad \text{for } P-a.e.\ \omega \in \Omega, \tag{2}$$

as well as for the mean scores that

$$E\big(S(\hat{Y},Y)\big) \leq E\big(S(Z,Y)\big) \tag{3}$$

with equality in (2) or (3) if and only if $\hat{Y} = Z$ $P-a.s.$

Let us turn to the situation where forecasts can be issued on the basis of two distinct information sets $\mathcal{F} \subset \mathcal{G} \subset \mathcal{A}$. Evidently, the larger information set should only yield better ideal forecasts, and indeed, we have the following result.

COROLLARY 2. *Suppose that $\mathcal{F} \subset \mathcal{G} \subset \mathcal{A}$ are increasing information sets. Set*

$$\hat{Y}_{\mathcal{F}}(\omega) = T\big(F_{Y|\mathcal{F}}(\omega,\cdot)\big), \qquad \hat{Y}_{\mathcal{G}}(\omega) = T\big(F_{Y|\mathcal{G}}(\omega,\cdot)\big). \tag{4}$$

Then

$$\begin{aligned} E\big(S(\hat{Y}_{\mathcal{G}},Y)|\mathcal{G}\big) &\leq E\big(S(\hat{Y}_{\mathcal{F}},Y)|\mathcal{G}\big) \qquad P-a.s.,\\ E\big(S(\hat{Y}_{\mathcal{G}},Y)|\mathcal{F}\big) &\leq E\big(S(\hat{Y}_{\mathcal{F}},Y)|\mathcal{F}\big) \qquad P-a.s.,\\ E\big(S(\hat{Y}_{\mathcal{G}},Y)\big) &\leq E\big(S(\hat{Y}_{\mathcal{F}},Y)\big), \end{aligned} \tag{5}$$

with equality in any of the inequalities in (5) if and only if $\hat{Y}_{\mathcal{F}} = \hat{Y}_{\mathcal{G}}$ $P-a.s.$

Thus, increasing the information set always leads to better ideal forecasts in terms of the score, except if the smaller information set already gives the same forecasts for the corresponding functional.
Finally, we point out that the equality $\hat{Y}_{\mathcal{F}} = \hat{Y}_{\mathcal{G}}$ $P-a.s.$ does not imply that the conditional distributions are equal, as the following example shows.

Example 1. We give an example involving quantiles. For a strictly increasing, continuous distribution function F let $q_\alpha(F)$ denote the α quantile, $\alpha \in (0,1)$, and let q_α be the α quantile of the standard normal distribution $N(0,1)$. Fix $\alpha \in (0,1)$, $\sigma > 1$, and let B, X_1, X_2 be independent random variables with $B \sim Ber(1/2)$, $X_1 \sim N(0,1)$, $X_2 \sim N\big(q_\alpha(1-\sigma), \sigma^2\big)$, and set $Y = BX_1 + (1-B)X_2$. If $\mathcal{F} = \{\emptyset, \Omega\}$ is trivial and $\mathcal{G} = \sigma\{B\}$, then the conditional distributions of Y are

$$\mathcal{L}(Y|\mathcal{F}) = \frac{1}{2}\,N(0,1) + \frac{1}{2}\,N\big(q_\alpha(1-\sigma), \sigma^2\big), \qquad \mathcal{L}(Y|\mathcal{G}) = B\,N(0,1) + (1-B)\,N\big(q_\alpha(1-\sigma), \sigma^2\big)$$

and in both cases, the conditional α quantile is constant and equals q_α.

Indeed, in order to evaluate the complete forecast distribution, strictly proper scoring rules are needed, as discussed in the next section.

2.3. *Probabilistic forecasts and proper scoring rules.* Let us briefly discuss general proper scoring rules, see Gneiting and Raftery (2007) for a detailed exposition. Recall that we identify the distribution functions $F \in \Theta$ with their associated probability measures $\mu_F \in \Theta$. A measurable mapping $\boldsymbol{S} : \Theta \times D \to \mathbb{R}$ is called a *scoring rule*. It is called *proper*, if for any $\mu \in \Theta$,

$$E_\mu\big(\boldsymbol{S}(\mu, Y)\big) \le E_\mu\big(\boldsymbol{S}(\nu, Y)\big) \qquad \text{for all } \nu \in \Theta, \tag{6}$$

and *strictly proper* if there is equality in (6) if and only if $\mu = \nu$. Gneiting (2011) points out that a functional T together with a consistent scoring function S induces the proper scoring rule $\boldsymbol{S}(\mu_F, y) = S\big(T(F), y\big)$. However, even if S is strictly consistent, $\boldsymbol{S}$ will not necessarily be strictly proper.
Let again $(\Omega, \mathcal{A}, P)$ be a probability space, and let $\mathcal{F} \subset \mathcal{A}$ be a sub-σ-algebra of $\mathcal{A}$ (the information set). A *Markov kernel* $G_{\mathcal{F}}$ (from $(\Omega, \mathcal{F})$ to $(\mathbb{R}, \mathcal{B})$) is a mapping

$$G_{\mathcal{F}} : \Omega \times \mathcal{B} \to [0,1],$$

such that
1. for any $\omega \in \Omega$, $B \mapsto G_{\mathcal{F}}(\omega, B)$ $(B \in \mathcal{B})$ is a probability measure on $(\mathbb{R}, \mathcal{B})$,
2. for any $B \in \mathcal{B}$, $\omega \mapsto G_{\mathcal{F}}(\omega, B)$ is $\mathcal{F} - \mathcal{B}[0,1]$-measurable.

The *(regular) conditional distribution* $\mu_{Y|\mathcal{F}}$ of Y given $\mathcal{F}$ is a particular Markov kernel (from $(\Omega, \mathcal{F})$ to $(\mathbb{R}, \mathcal{B})$) such that for all $B \in \mathcal{B}$,

$$E\big(1_{Y \in B}|\mathcal{F}\big)(\omega) = \mu_{Y|\mathcal{F}}(\omega, B) \qquad \text{for } P-a.e.\ \omega \in \Omega.$$

Theorem 3. *Let $\boldsymbol{S}$ be a strictly proper scoring rule. Let $\mu_{Y|\mathcal{F}}(\omega, \cdot)$ be the conditional distribution of Y given $\mathcal{F}$. Assume that for each ω, $\mu_{Y|\mathcal{F}}(\omega, \cdot) \in \Theta$. For any Markov kernel $G_{\mathcal{F}}$*

(from $(\Omega, \mathcal{F})$ to $(\mathbb{R}, \mathcal{B})$) for which $G_{\mathcal{F}}(\omega, \cdot) \in \Theta$ for all $\omega \in \Omega$, the map $\omega \mapsto \boldsymbol{S}\big(G_{\mathcal{F}}(\omega, \cdot), Y(\omega)\big)$ is a random variable and we have that

$$E\left(\boldsymbol{S}(\mu_{Y|\mathcal{F}}, Y)|\mathcal{F}\right)(\omega) \leq E\left(\boldsymbol{S}(G_{\mathcal{F}}, Y)|\mathcal{F}\right)(\omega) \qquad \text{for } P-a.e.\ \omega \in \Omega \tag{7}$$

and

$$E\left(\boldsymbol{S}(\mu_{Y|\mathcal{F}}, Y)\right) \leq E\left(\boldsymbol{S}(G_{\mathcal{F}}, Y)\right) \tag{8}$$

with equality in (7) or (8) if and only if for $P-a.e.\ \omega \in \Omega$, the distributions $G_{\mathcal{F}}(\omega, \cdot)$ and $\mu_{Y|\mathcal{F}}(\omega, \cdot)$ coincide.

This is also observed in Tsyplakov (2011) in his comment on the paper by Mitchell and Wallis (2011) which in turn was a critical response to Gneiting, Balabdaoui and Raftery (2007). Gneiting, Balabdaoui and Raftery (2007) discuss the somewhat too dominant role of the probability integral transform (PIT) in evaluating forecasts. They focus on the uniformity of the PIT if the forecasts are correctly specified. Tsyplakov (2011) also indicates a result similar to Proposition 6 (see Section 3.1) for the PIT and observes that mere independence of the PIT values does not adequately take into account the role of the information set.

Corollary 4. *Let $\mathcal{F} \subset \mathcal{G} \subset \mathcal{A}$ be increasing information sets. If $\boldsymbol{S}$ is a strictly proper scoring rule and for each ω, $\mu_{Y|\mathcal{F}}(\omega, \cdot), \mu_{Y|\mathcal{G}}(\omega, \cdot) \in \Theta$, then*

$$E\left(\boldsymbol{S}(\mu_{Y|\mathcal{G}}, Y)|\mathcal{G}\right)(\omega) \leq E\left(\boldsymbol{S}(\mu_{Y|\mathcal{F}}, Y)|\mathcal{G}\right)(\omega) \qquad \text{for } P-a.e.\ \omega \in \Omega \tag{9}$$

and

$$E\left(\boldsymbol{S}(\mu_{Y|\mathcal{G}}, Y)\right) \leq E\left(\boldsymbol{S}(\mu_{Y|\mathcal{F}}, Y)\right), \tag{10}$$

with equality in (9) or (10) if and only if for $P-a.e.\ \omega \in \Omega$, the conditional distributions $\mu_{Y|\mathcal{G}}(\omega, \cdot)$ and $\mu_{Y|\mathcal{F}}(\omega, \cdot)$ coincide.
If in particular $\mathcal{G} = \sigma(\mathcal{F}, \mathcal{H})$, where $\mathcal{H} \subset \mathcal{A}$ is another sub-σ-algebra, then there is equality in (9) or (10) if and only if Y and $\mathcal{H}$ are conditionally independent given $\mathcal{F}$.

Thus, using a strictly proper scoring rule to evaluate the complete predictive distribution, the predictive distributions in Example 1 based on distinct information sets could be distinguished. However, if interest is focused on a single functional like the mean or the VaR, then this might not be necessary. The second part of the corollary extends results by Bröcker (2009) and DeGroot and Fienberg (1983) from finite to general real state space.

2.4. *Testing for sufficient information.* Consider the setting of Section 2.2 in which the aim is to forecast a functional $T : \Theta \to \mathbb{R}$. When evaluating forecasts empirically, one observes a sequence of forecasts $\hat{Y}_1, \ldots, \hat{Y}_N$ of T with the corresponding realizations $Y_1, \ldots, Y_N$, and proceeds by averaging the corresponding scores.
More specifically, assume that $(Y_n)_{n\geq 1}$ is a stationary and ergodic sequence, and let $(\mathcal{F}_n)_{n\geq 1}$ be a filtration (increasing sequence of sub-σ-algebras of $\mathcal{A}$) such that Y_n is $\mathcal{F}_n$-measurable, $n \geq 1$. Suppose that the h-step forecasts

$$\hat{Y}^{(h)}_{n,\mathcal{F}}(\omega) := \hat{Y}_{\mathcal{F}_{n-h}}(\omega) = T\big(F_{Y_n|\mathcal{F}_{n-h}}(\omega, \cdot)\big)$$

are stationary and ergodic as well. Then for the averaged loss, as $N \to \infty$,

$$\widehat{m}_{N,\mathcal{F}} := \frac{1}{N}\sum_{n=1}^{N} S(\hat{Y}_{n,\mathcal{F}}^{(h)}, Y_n) \to E\big(S(\hat{Y}_{1,\mathcal{F}}^{(h)}, Y_1)\big) \qquad P-a.s. \tag{11}$$

In this section we investigate the behaviour of the classical Diebold and Mariano (1995) test when evaluating asymptotically ideal forecasts based on distinct, nested information sets using strictly consistent scoring functions. See below for further discussion on the relation to the literature.

Suppose that $(\mathcal{G}_n)$ is a second filtration for which $\mathcal{F}_n \subset \mathcal{G}_n$ for all $n \geq 1$, and for which the sequence $\hat{Y}_{n,\mathcal{G}}^{(h)} := \hat{Y}_{\mathcal{G}_{n-h}}$ is stationary and ergodic as well. We shall propose a test for the hypothesis

$$H : \hat{Y}_{n,\mathcal{G}}^{(h)} = \hat{Y}_{n,\mathcal{F}}^{(h)} \qquad P-a.s. \qquad \text{for all } n \geq 1, \tag{12}$$

that both sequences of information sets lead to the same forecasts. By stationarity, this is equivalent to $\hat{Y}_{1,\mathcal{G}}^{(h)} = \hat{Y}_{1,\mathcal{F}}^{(h)}$ $P-a.s.$.

The h-step forecasts for time n based on $\mathcal{G}_{n-h}$ and on $\mathcal{F}_{n-h}$ which are actually issued are denoted by $\tilde{Y}_{n,\mathcal{G}}^{(h)}$ and $\tilde{Y}_{n,\mathcal{F}}^{(h)}$. Since we are concerned with the ideal forecasts, we need to make the rather strong assumption that the errors (due to misspecification and estimation effects) in these sequences of forecasts have asymptotically negligible effect on the scores. More precisely, consider the following conditions.

$$\sum_{n=1}^{N}\Big(S(\tilde{Y}_{n,\mathcal{J}}^{(h)}, Y_n) - S(\hat{Y}_{n,\mathcal{J}}^{(h)}, Y_n)\Big) = o_P(\sqrt{N}) \quad \big(\text{or} \quad = O_P(\sqrt{N})\big), \qquad \mathcal{J} = \mathcal{F}, \mathcal{G}. \tag{13}$$

As a test statistic, consider

$$M_N = \frac{1}{N}\sum_{n=1}^{N}\Big(S(\tilde{Y}_{n,\mathcal{F}}^{(h)}, Y_n) - S(\tilde{Y}_{n,\mathcal{G}}^{(h)}, Y_n)\Big) = \widehat{m}_{N,\mathcal{F}} - \widehat{m}_{N,\mathcal{G}}.$$

Theorem 5. *Under the above stationarity assumptions suppose that* $E\big(S(\hat{Y}_{1,\mathcal{F}}^{(h)}, Y_1)^2\big) < \infty$ *is satisfied. Under the null hypothesis* H *in (12), if (13) holds with* $o_P(\sqrt{N})$, *then*

$$\sqrt{N}\, M_N \xrightarrow{d} N\big(0, \sigma^2\big),$$
$$\sigma^2 = E\left(Z_1^2 + 2\sum_{n=2}^{h} Z_1 Z_n\right), \qquad Z_n = S\big(\hat{Y}_{n,\mathcal{F}}^{(h)}, Y_n\big) - S\big(\hat{Y}_{n,\mathcal{G}}^{(h)}, Y_n\big). \tag{14}$$

Under an alternative, if (13) holds with $O_P\big(\sqrt{N}\big)$, *we get* $\sqrt{N} M_N \to \infty$ *in probability.*

Let us give some remarks on the above result.

1. Suppose that $\hat{\sigma}_N^2$ is a consistent estimate of the long run variance σ^2. Then form the t-statistic

$$T_N = \sqrt{N}\, M_N / \hat{\sigma}_N,$$

which under the hypothesis H is asymptotically $N(0,1)$-distributed. One chooses a one-sided rejection region and rejects with asymptotic level α if $T_N > q_{1-\alpha}$. If under the alternative, $\hat{\sigma}_N$ remains bounded, we obtain $T_N \to \infty$ in probability, so that the test is consistent.

Estimation of the long-run variance σ^2 is a delicate task. There is a large literature starting with Newey and West (1987), who already propose weights in (14) which guarantee non-negativity as well as consistency. In our situation, one could truncate the series at the fixed prediction window h and use weights one. While this works under the hypothesis, in our simulations a higher value of $2h$ for the truncation with constant weights of value one gave better power properties. Further, since under the alternative the observations do not have mean zero, we computed actual covariances including centering (not just second moments).

2. There is a huge econometric literature on comparing the predictive accuracy of competing forecasts, starting with the classic paper by Diebold and Mariano (1995). For a sequence of forecasts, $\hat{y}_1, \ldots, \hat{y}_N$, and corresponding observations $y_1, \ldots, y_N$, typically the forecast errors $e_n = y_n - \hat{y}_n$ are formed, and these are inserted into a certain loss function $l(e)$. For a competing sequence of forecasts, $\hat{z}_1, \ldots, \hat{z}_N$, the same process is applied, leading to $\tilde{e}_n = y_n - \hat{z}_n$. The Diebold and Mariano (DM) test statistic is now based on analyzing the asymptotic distribution of

$$\tilde{M}_N = \frac{1}{N} \sum_{n=1}^{N} \big(l(e_n) - l(\tilde{e}_n) \big).$$

Under stationarity assumptions on the sequences of errors (e_n) and $(\tilde{e}_n)$, the asymptotic distribution of $\tilde{M}_N$ may be analyzed, and a t-statistic with a two-sided rejection region may be formed.

We note that if the forecasts $\hat{y}_n$ correspond to a certain functional T and a sequence of information sets, and if the scoring function S is a function in the difference e_n, then our test is simply the DM test, and we analyze its behaviour for asymptotically ideal forecasts based on two distinct, ordered information sets.

As a conclusion, in our situation the DM test, performed as a one-sided test, is a consistent test for testing the effect of increasing the information set on the forecast of the functional. Of course, the assumption of asymptotically ideal forecasts is a strong one. However, if it does not hold, the test remains valid as long as the observations and both sequences for forecasts remain stationary and the CLT still applies (see Durrett, 2005, p. 416, Theorem (7.6), for sufficient conditions), in the sense that it keeps its asymptotic level (of course, the test is then no longer consistent).

The point of view of relating the loss function precisely to the functional to be predicted is usually not pursued in the econometric literature (but see Gneiting and Ranjan, 2011), which is often not particularly precise on what (meaning which functional of the predictive distribution) is actually forecast. Using the "wrong" loss function for a specific functional may result in strongly biased results, see the example in Section 1.2 in Gneiting (2011). Further, there are scoring functions which are not functions in the linear forecast errors $e = y - \hat{y}$, cf. Gneiting (2011).

Diebold (2012) revisits the DM test, and in particular points out distinctions between comparing forecasting models (forecasts arising from specific econometric models), forecasting methods (from models but taking into account the effect of parameter estimation, see e.g. Giacomini and White, 2006), or mere forecasts like in the DM test, no matter how these were generated. Our approach is well in line with Diebold as we simply compare forecasts. If these are (at least asymptotically) ideal, the effect of increasing the information set on the functional may be tested consistently.

3. We conclude this subsection by remarking that the above test may be extended to the case of proper scoring rules, in order to evaluate the effect of increasing the information set on the

complete predictive distribution.

3. Backtesting value at risk estimates. The most widely used risk measure in quantitative finance is the value at risk (VaR), see e.g. Jorion (2006), Christoffersen (2009) or McNeil, Frey and Embrechts (2005). Formally, this is a (high, say 0.99 or 0.999) quantile of the loss distribution.

For issuing VaR forecasts, different variations exist. Unconditional methods base the VaR on the unconditional distribution of the risk factors thus using a trivial information set, while conditional methods refer to a conditional distribution typically given the historical data. Here, the information set may vary as well, in a portfolio point of view it only includes the portfolio returns, while a modeling of the individual risk factors involves a larger information set. See also Section 4 for further details.

Following Berkowitz, Christoffersen and Pelletier (2011), typical areas of application of VaR estimates include:

A. Risk controls for trading desks. The distinct trading desks (equities, currencies, derivatives, fixed-income) have limits for the VaR, typically one-day ahead, of their trading position. These are set by the management and monitored in real-time by the back-office.

B. Portfolio choice. Instead of the classical Markowitz mean-variance portfolio optimization, the VaR is used as a risk measure when forming the optimal portfolios. Here, longer time horizons (month, quarter) are considered, and a multivariate modeling of the risk factors is required, see Christoffersen (2009).

C. Regulatory uses. Commercial banks are required to hold a certain amount of safe assets. When based on internal methods, this amount is determined as a function of the VaR, over a two-week horizon, and at a level of 99%.

Different goals may be pursued for the specific VaRs reported in each scenario. For example, in case C., the bank will be interested to report a "small" (but still valid) VaR so that the required amount of regulatory capital is reasonably small. Further, the VaR reported in C. should not vary too much over time, since the regulatory capital can and should not be shifted abruptly.

In any of the three cases, it is of major interest to quantify and minimize the expected amount of losses resulting from exceedences of the VaR estimates which are being reported. To this end, our result which relates the expected score for the VaR to the expected shortfall is of major interest. Below we deduce from Corollary 2 that ideal VaR forecasts are improved in terms of the expected shortfall arising from their exceedences by increasing the information set.

3.1. *Exceedance indicators.* Evaluating the VaR forecasts is called backtesting. In unconditional backtesting, one checks whether the relative frequency of exceedances of the VaR estimates corresponds to the level of the VaR, see McNeil, Frey and Embrechts (2005). While both unconditional and conditional methods (if correctly specified) keep the level, the empirical level of exceedances alone does not imply that the sequence of forecasts issued is actually related to a quantile. Indeed, suppose that $\alpha = 0.99$, then simply issue systematically 99 extremely high values followed by a single extremely low value (resulting in non-stationary forecasts). This way, a very quick convergence of the empirical exceedances to the nominal level will be observed, but the forecasts do not make sense.

Conditional methods are often accompanied by independence checks, the basis of which is the following well-known proposition. For a strictly increasing, continuous distribution function

F let $q_\alpha(F)$ denote the α quantile.

PROPOSITION 6. *Suppose that for each* $\omega \in \Omega$*, the conditional distribution function* $F_{Y|\mathcal{F}}(\omega, \cdot)$ *is continuous and strictly increasing. Let* Z *be an* $\mathcal{F}$*-measurable random variable and let* $I = 1_{Y>Z}$ *be the exceedance indicator. Then the following assertions 1. and 2. are equivalent:*
1. $P(I = 1) = 1 - \alpha$*, and* I *and* $\mathcal{F}$ *are independent.*
2. $Z(\omega) = q_\alpha\big(F_{Y|\mathcal{F}}(\omega, \cdot)\big)$ *for* $P-a.e.$ $\omega \in \Omega$*.*

The proposition implies the following so-called i.i.d. and hence the joint hypothesis (see Christoffersen, 1998).

COROLLARY 7. *Suppose that* (Y_n) *is a sequence of random variables and that* $(\mathcal{F}_n)$ *is any filtration to which* (Y_n) *is adapted (i.e.* Y_n *is* $\mathcal{F}_n$*-measurable). Suppose further that all conditional distribution functions* $F_{Y_n|\mathcal{F}_{n-1}}$ *are continuous and strictly increasing. Then for the one-step prediction* $\hat{Y}_n = q_\alpha\big(F_{Y_n|\mathcal{F}_{n-1}}(\omega, \cdot)\big)$*, the sequence of exceedance indicators* $I_n = 1_{Y_n > \hat{Y}_n}$ *is independent and Bernoulli distributed with success probability* $1 - \alpha$*.*

Some remarks are in order.
1. The corollary is useful for checking whether for a given sequence of information sets, a certain forecasting method which will be based on specification and testing works adequately. Several tests have been proposed, taking into account effects of model misspecification and estimation schemes, cf. Escanciano and Olmo (2011).
2. However, as remarked e.g. in Escanciano and Olmo (2011), mere independence of the exceedance indicators does not appropriately take into account the role of the sequence of information sets $(\mathcal{F}_n)$, since all that is needed is that (Y_n) is adapted to $(\mathcal{F}_n)$.
3. When increasing the information sets, e.g. by multivariate modeling of risk factors, one cannot expect that the average of the exceedance indicators will be systematically closer to the level $1 - \alpha$, which is, however, sometimes taken as a criterion (see McNeil, Frey and Embrechts, 2005, pp. 55–59). Indeed, the speed of convergence in $\frac{1}{N}\sum_{n=1}^N I_n \to 1 - \alpha$ for independent (I_n) is governed by the central limit theorem

$$\sqrt{N}\left(\frac{1}{N}\sum_{n=1}^{N} I_n - (1-\alpha)\right) \xrightarrow{d} N\big(0, \alpha(1-\alpha)\big).$$

In order to decrease the asymptotic variance $\alpha(1-\alpha)$ negatively correlated exceedance indicators are required, and in order to attain a faster rate than $\sqrt{N}$, nonstationary forecasts need to be issued as in the stylized example above.
4. The situation is even worse for h-step forecasts, which are therefore comparatively rarely investigated in academic studies. Here, $\hat{Y}_n^{(h)} = q_\alpha\big(F_{Y_n|\mathcal{F}_{n-h}}(\omega, \cdot)\big)$, and exceedance indicators $I_n = 1_{Y_n > \hat{Y}_n^{(h)}}$ are only independent for lags $\geq h$.
5. In principle, the VaR based on the specific information set $\mathcal{F}_{n-h}$ can be identified from the exceedance indicator by checking full independence against the information set $\mathcal{F}_{n-h}$, see Proposition 6. Some tests take into account the required independence of exceedance indicators to additional lagged variables, see Berkowitz, Christoffersen and Pelletier (2011). However, the question arises what the particular additional gain is from this extended independence property.

3.2. *Quantile loss and the expected shortfall.* In what sense are ideal VaR forecasts then improved by increasing the information set? A suitable answer seems to be provided by the theory of the previous section, using scoring functions.
Indeed, the α quantile is elicitable, and the strictly consistent scoring functions satisfying (1) are given by

$$S(x,y) = \big(1_{x\geq y} - \alpha\big)\,\big(g(x) - g(y)\big), \tag{15}$$

where g is strictly increasing (and all relevant expected values are assumed to exist), see Gneiting (2011). Note that we can drop the term $\alpha g(y)$ from (15) and retain a strictly consistent scoring function (though no longer non-negative, and not necessarily satisfying (1)). An attractive special case is the choice $g(x) = x/\alpha$. After substracting y, we arrive at the (no longer non-negative) strictly consistent scoring function

$$S^*(x,y) = \frac{1}{\alpha} 1_{x\geq y}\big(x - y\big) - x = x\,\big(\alpha^{-1} 1_{x\geq y} - 1\big) - y\,\alpha^{-1}\,1_{x\geq y}.$$

Now we relate the score under S^* to the expected shortfall.

Proposition 8. *Suppose that Y is integrable and that for each $\omega \in \Omega$, the conditional distribution function $F_{Y|\mathcal{F}}(\omega,\cdot)$ is continuous and strictly increasing. For the conditional quantile $\hat{Y}_{\mathcal{F}}(\omega) = q_\alpha\big(F_{Y|\mathcal{F}}(\omega,\cdot)\big)$ we get*

$$E\Big(S^*\big(\hat{Y}_{\mathcal{F}}, Y\big)|\mathcal{F}\Big)(\omega) = -\frac{1}{\alpha}\int_{-\infty}^{\hat{Y}_{\mathcal{F}}(\omega)} y\, F_{Y|\mathcal{F}}(\omega, dy) \qquad \text{for } P - a.e.\ \omega \in \Omega. \tag{16}$$

Moreover, if $\mathcal{F} \subset \mathcal{G} \subset \mathcal{A}$ and $\hat{Y}_{\mathcal{G}}(\omega) = q_\alpha\big(F_{Y|\mathcal{G}}(\omega,\cdot)\big)$, then

(17)
$$\begin{aligned}
-\frac{1}{\alpha}\int_{-\infty}^{\hat{Y}_{\mathcal{G}}(\omega)} y\, F_{Y|\mathcal{G}}(\omega, dy) &\leq -\frac{1}{\alpha}\int_{-\infty}^{\hat{Y}_{\mathcal{F}}(\omega)} y\, F_{Y|\mathcal{F}}(\omega, dy) \qquad \text{for } P - a.e.\ \omega \in \Omega,\\
E\Big(-\frac{1}{\alpha}\int_{-\infty}^{\hat{Y}_{\mathcal{G}}(\cdot)} y\, F_{Y|\mathcal{G}}(\cdot, dy)\,\Big|\mathcal{F}\Big)(\omega) &\leq -\frac{1}{\alpha}\int_{-\infty}^{\hat{Y}_{\mathcal{F}}(\omega)} y\, F_{Y|\mathcal{F}}(\omega, dy) \qquad \text{for } P - a.e.\ \omega \in \Omega,\\
\int_\Omega -\frac{1}{\alpha}\int_{-\infty}^{\hat{Y}_{\mathcal{G}}(\omega)} y\, F_{Y|\mathcal{G}}(\omega, dy)\, dP(\omega) &\leq \int_\Omega -\frac{1}{\alpha}\int_{-\infty}^{\hat{Y}_{\mathcal{F}}(\omega)} y\, F_{Y|\mathcal{F}}(\omega, dy)\, dP(\omega).
\end{aligned}$$

with equality in one of the inequalities in (17) if and only if $\hat{Y}_{\mathcal{G}} = \hat{Y}_{\mathcal{F}}$ a.s.

For an interpretation, suppose that Y corresponds to the profit and loss distribution (e.g. is a log-return), so that α is indeed a small value such as $\alpha = 0.01$ or 0.001. Then

$$-\frac{1}{\alpha}\int_{-\infty}^{\hat{Y}_{\mathcal{F}}(\omega)} y\, F_{Y|\mathcal{F}}(\omega, dy)$$

is the lower-tail expected shortfall of the conditional distribution, and thus $E\big(S^*(\hat{Y}_{\mathcal{F}}(\omega), Y)\big)$ as in (17) is the mean lower-tail expected shortfall when using the information set $\mathcal{F}$. Rockafellar and Uryasev (2000) give a result similar to (16), see their Theorem 1.

4. Simulations and Applications. In this section we investigate the proposed methods in the context of value at risk estimation both in simulated examples as well as for log-returns of several stocks and stock indices. We let $T : \Theta \to \mathbb{R}$ be the α quantile, and let $S(x, y) = x\left(\alpha^{-1} 1_{x \geq y} - 1\right) - y\, \alpha^{-1} 1_{x \geq y}$, see Section 3. While the quantile loss function has been used in some numerical studies (c.f. Bao, Lee and Saltoğlu, 2006), the particular effect of the information set does not seem to have been investigated so far.

4.1. *Unconditional versus conditional risk management.* We consider the situation of conditional versus unconditional risk management, see McNeil, Frey and Embrechts (2005). Let $(R_t)_{t \in \mathbb{Z}}$ be a stationary time series corresponding to daily log-returns of a stock or stock index, and let

$$\mathcal{F}_t = \{\emptyset, \Omega\}, \qquad \mathcal{G}_t = \sigma\{R_s \colon s \leq t\}.$$

Thus, forecasts based on the trivial $\mathcal{F}_t$ concern the unconditional distribution of returns, while forecasts based on $\mathcal{G}_t$ concern the conditional distribution given daily log-returns. Fix some prediction horizon $h \geq 1$, and set

$$Y_{t+h} = Y_{t+h}^{(h)} = R_{t+1} + \ldots + R_{t+h},$$

the h-step log-return. Our aim is h-step forecasting of the quantile of Y_t, i.e.

$$\hat{Y}_{t+h,\mathcal{F}}^{(h)} = T\big(F_{Y_{t+h}|\mathcal{F}_t}\big) \qquad \text{and} \qquad \hat{Y}_{t+h,\mathcal{G}}^{(h)} = T\big(F_{Y_{t+h}|\mathcal{G}_t}\big).$$

Since $\mathcal{F}_t$ is trivial, the $\hat{Y}_{t+h,\mathcal{F}}^{(h)}$ are constant a.s. and equal to the unconditional quantile of the Y_t, while $\hat{Y}_{t+h,\mathcal{G}}^{(h)}$ is the conditional quantile of the h-step return given the history of one-step returns up to time t. For the conditional method, the exceedance indicators $1_{\hat{Y}_{t+h,\mathcal{G}}^{(h)} > Y_{t+h}}$ are independent for lags $\geq h$, while there is no such general independence for the unconditional method. However, note that for larger values of h, it is quite hard to distinguish both methods based on (non) independence.

Simulation

As data-generating process, we use a GARCH$(1, 1)$-model

$$R_t = \sigma_t \epsilon_t, \qquad \sigma_t^2 = \kappa + \phi R_{t-1}^2 + \beta \sigma_{t-1}^2,$$

where the (ϵ_t) are i.i.d. $N(0, 1)$-distributed, and the distinct parameter values (κ, ϕ, β) are chosen according to the scenarios in Table 1. As prediction horizons we consider $h = 1, 2$: one and two days, $h = 10$: two weeks, $h = 66$: one quarter of the year. Given the parameters of the GARCH model (either true values or estimates) as well as estimates of the one-step volatilities σ_t^2, as conditional forecasts we use in case $h = 1$ the exact forecast distribution $N(0, \sigma_{t+1}^2)$, while for $h > 1$ we approximate the quantile by the empirical quantile of a Monte-Carlo sample of size $M = 1000$ for each t. As unconditional forecast we use an α quantile of an appropriate series of h-step returns.

a. First, we briefly investigate the true expected mean scores for unconditional and conditional risk management using (approximate) ideal forecasts, which by (17) correspond to average expected shortfalls. To this end, we use a single huge sample of size $N = 100{,}000$ (resp. $N = 300{,}000$ for $h = 1$). For the conditional forecasts $\hat{Y}_{t+h,\mathcal{G}}^{(h)}$, we use the true parameters of the GARCH model, while for the unconditional case, we set $\hat{Y}_{t+h,\mathcal{F}}^{(h)}$ constant as the empirical quantile of a distinct simulated series of (Y_t) of length 300,000. Finally, we approximate the mean score by the sample averages $\widehat{m}_{N,\mathcal{F}}$ and $\widehat{m}_{N,\mathcal{G}}$ as in (11).

The results for configuration 1 can be found in Table 2, for the other configurations these are similar. As stated in Acerbi and Tasche (2002, Proposition 3.4), we see that for fixed h and increasing values of α, the values of $\widehat{m}_{N,\mathcal{F}}$ and $\widehat{m}_{N,\mathcal{G}}$ decrease. Moreover, for fixed α and increasing values of h, $\widehat{m}_{N,\mathcal{F}}$ and $\widehat{m}_{N,\mathcal{G}}$ increase. The relative difference, $M_N/\hat{\mu}_{N,\mathcal{F}}$, which indicates the reduction in mean expected shortfall when passing from the unconditional to the conditional method, is highest for small α for fixed h, with values as large as 31%.

The estimate $\hat{\sigma}^2$ for $\sigma^2 = E(Z_1^2 + 2\sum_{k=2}^{\infty} Z_1 Z_k)$, where the Z_k are as in (14), is obtained by truncation at $2h$ with constant weight one, and where the observations are centered before computing covariances. This choice gave reasonable power properties in our simulations.

The last column contains the values T_N of the t-statistic together with the p-value based on the asymptotic approximation. For the values $h = 1, 2$ and 10, the difference is significantly > 0 for all α, while for $h = 66$ it is not significant.

Config	κ	ϕ	β
1	0.01	0.088	0.902
2	0.02	0.2	0.78
3	0.05	0.3	0.65

TABLE 1

Parameter configurations for the GARCH(1,1) model in the comparison of conditional versus unconditional VaR estimation

h	α	Mean Scores $\hat{m}_{N,\mathcal{F}}$	$\hat{m}_{N,\mathcal{G}}$	Diff. ($= M_N$) $\hat{m}_{N,\mathcal{F}} - \hat{m}_{N,\mathcal{G}}$	Rel. diff. $M_N/\hat{m}_{N,\mathcal{F}}$	$\hat{\sigma}$	T_N	$\Pr(> T_N)$
1	0.01	3.627	2.511	1.116	0.31	15.7	38.9	< 0.001
1	0.05	2.225	1.895	0.330	0.15	3.4	52.4	< 0.001
1	0.20	1.354	1.303	0.051	0.04	0.7	40.5	< 0.001
2	0.01	4.547	3.573	0.974	0.21	18.9	8.1	< 0.001
2	0.05	3.652	3.035	0.617	0.17	8.0	12.3	< 0.001
2	0.20	1.882	1.828	0.055	0.03	1.2	7.5	< 0.001
10	0.01	12.852	9.579	3.272	0.25	115.8	4.5	< 0.001
10	0.05	6.749	5.988	0.761	0.11	19.5	6.2	< 0.001
10	0.20	3.991	3.890	0.101	0.03	4.6	3.5	< 0.001
66	0.01	25.331	24.746	0.585	0.02	320.9	0.3	0.387
66	0.05	17.726	17.398	0.328	0.02	76.4	0.7	0.249
66	0.20	11.552	11.407	0.145	0.01	28.2	0.8	0.209

TABLE 2

Mean scores for conditional and unconditional VaR estimation for parameter configuration 1, see Table 1

b. Next, we investigate the power of the resulting DM test for realistic sample sizes when taking into account estimation effects. We based estimation of the parameters of the GARCH model for the unconditional method as well as of the quantile for the unconditional method on a rolling window of size $R_{\text{wind}} = 500$. For the unconditional method, we investigated two variations, first using the empirical quantile of the last R_{wind} h-step returns preceding t, and second using a square root of time rule resulting in $\hat{Y}_{t+h}^{(h)} = \sqrt{h}\hat{s}_t q_\alpha + h\hat{m}_t$, where $\hat{s}_t$ and $\hat{m}_t$ are the empirical standard deviation and mean of the last R_{wind} one-step returns preceding

		Config		
h	N	1	2	3
1	250	0.463	0.565	0.479
	500	0.632	0.776	0.640
	1000	0.863	0.951	0.900
	1500	0.947	0.993	0.981
2	250	0.392	0.421	0.326
	500	0.492	0.576	0.447
	1000	0.723	0.844	0.744
	1500	0.859	0.957	0.905
	2000	0.920	0.984	0.970
	4000	0.999	0.999	0.999
10	250	0.258	0.214	0.140
	500	0.196	0.157	0.087
	1000	0.205	0.173	0.079
	1500	0.277	0.232	0.119
	2000	0.330	0.306	0.162
	4000	0.634	0.620	0.412

TABLE 3

Power of the test (at the 0.05 *level) for conditional and unconditional VaR estimation* ($\alpha = 0.01$)*; for parameter configurations, cf. Table 1*

t, and q_α is the α quantile of the standard normal. Since the square root of time rule in most cases led to smaller scores, we only displayed the corresponding results. Note that due to the limited estimation horizon, the unconditional method is in fact also partially conditional. We then compute the DM t-statistic T_N with estimate for the long-run-variance as described above. This is iterated 1000 times.

Results for the three configurations of Table 1, various sample sizes N (so that the number of observations is $N + R_{\text{wind}}$), test levels 0.05 and 0.1 and $h = 1, 2, 10$ are displayed in Tables 3 and 4. For $h = 66$, the test does not have any power beyond the level. Otherwise, the power properties are quite reasonable.

Application

Finally we investigate unconditional versus conditional risk management when applied to log-returns of several stocks and stock-indices. We use publicly available share prices of German stocks (on a daily basis) from Yahoo Finance (`http://finance.yahoo.com`). The data set runs from 1st January 2001 to 31st July 2013. In the direct comparison of two shares we restrict for simplicity to the subset of available data points (for each share) by taking intersections. In any case, the subset of share prices in our analysis was larger than 2727 (each including the beginning of the year 2003). Let S_t denote the price, $R_t = \log S_t - \log S_{t-1}$ the log-return, so that

$$Y_{t+h}^{(h)} = \log S_{t+h} - \log S_t$$

is the h-step log-return. We proceed as in the simulations part b. above, using a rolling window of size 500 as well as the square-root of time rule for the unconditional method. The results for various stocks can be found in Table 5. The mean score is significantly reduced for $h = 1$ and $h = 2$ when passing from the unconditional to the conditional methods, where the maximal value for the relative difference is 0.3. For higher lags, the reduction is non-significant.

		Config		
h	N	1	2	3
1	250	0.578	0.693	0.629
	500	0.757	0.870	0.792
	1000	0.921	0.982	0.956
	1500	0.978	0.998	0.990
2	250	0.523	0.612	0.510
	500	0.666	0.766	0.688
	1000	0.858	0.937	0.912
	1500	0.934	0.987	0.981
	2000	0.974	0.995	0.991
	4000	1.000	1.000	1.000
10	250	0.350	0.288	0.206
	500	0.335	0.295	0.189
	1000	0.416	0.372	0.233
	1500	0.509	0.495	0.321
	2000	0.590	0.597	0.410
	4000	0.819	0.842	0.689

TABLE 4

Power of the test (at the 0.1 *level) for conditional and unconditional VaR estimation (*$\alpha = 0.01$*); for parameter configurations, cf. Table 1*

Conclusions

For $h = 1$ and $h = 2$, the improved performance of the conditional method compared to the unconditional method is apparent, both in the simulations and also in the stock returns. On the other hand, for $h = 66$ (the quarter) there is no significant improvement for the stock returns, and the potential improvement as indicated by the simulations is also small. For $h = 10$ (two weeks), simulations indicate quite a potential for improvement, but the effect in the actual stock returns, if present, is not yet significant.

4.2. *Univariate versus multivariate modeling for risk management.* Now we consider a univariate modeling on the basis of portfolio returns, versus a multivariate modeling of the individual risk factors. For simplicity we only investigate two underlying risk factors.
Let $(\mathbf{R}_t)_{t\in\mathbb{Z}}$, $\mathbf{R}_t = (R_{t,1}, R_{t,2})^T$ be a stationary bivariate time series corresponding to daily returns of the individual stocks of a portfolio. For a fixed weight vector $\mathbf{w} = (w_1, w_2)^T$, with $0 \le w_i \le 1$, $w_1 + w_2 = 1$, we let $Y_t = \mathbf{w}^T \mathbf{R}_t$, which we interpret as the return of a portfolio consisting of the two individual stocks. Note that on the basis of the prices of the portfolio, this corresponds to a reweighting in each step, see the application below. As information sets, consider

$$\mathcal{F}_t = \sigma\{Y_s \colon s \le t\}, \qquad \mathcal{G}_t = \sigma\{\mathbf{R}_s \colon s \le t\},$$

the history of portfolio returns $\mathcal{F}_t$ and of individual risk factors $\mathcal{G}_t$. Our aim is one-step forecasting of the quantile of Y_t, i.e.

$$\hat{Y}^{(1)}_{t+1,\mathcal{F}} = \hat{Y}_{t+1,\mathcal{F}} = T\big(F_{Y_{t+1}|\mathcal{F}_t}\big) \qquad \text{and} \qquad \hat{Y}^{(1)}_{t+1,\mathcal{G}} = \hat{Y}_{t+1,\mathcal{G}} = T\big(F_{Y_{t+1}|\mathcal{G}_t}\big).$$

Thus, $\hat{Y}_{t+1,\mathcal{F}}$ is the forecast based on the history of portfolio returns, while $\hat{Y}_{t+1,\mathcal{G}}$ is the forecast based on the history of individual risk factors. Note that in both cases, for ideal forecasts the series of exceedance indicators $(I_{t,\mathcal{F}})$ and $(I_{t,\mathcal{G}})$, where $I_{t,\mathcal{F}} = 1_{\hat{Y}_{t,\mathcal{F}} > Y_t}$ and $I_{t,\mathcal{G}} = 1_{\hat{Y}_{t,\mathcal{G}} > Y_t}$, are

Share Name	h	Mean Scores $\hat{m}_{N,\mathcal{F}}$	$\hat{m}_{N,\mathcal{G}}$	Diff. $(= M_N)$ $\hat{m}_{N,\mathcal{F}} - \hat{m}_{N,\mathcal{G}}$	Rel. diff. $M_N/\hat{m}_{N,\mathcal{F}}$	$\hat{\sigma}$	T_N	$\Pr(> T_N)$
DAX	1	5.87	4.24	1.63	0.28	15.4	5.48	0.000
	2	8.43	6.46	1.97	0.23	29.3	3.50	0.000
	10	22.07	18.26	3.80	0.17	67.8	2.91	0.002
Daimler	1	8.62	6.92	1.70	0.20	18.6	4.81	0.000
	2	12.14	9.75	2.39	0.20	36.6	3.42	0.000
	10	34.44	29.44	5.00	0.15	193.2	1.35	0.088
Deutsche Bank	1	10.08	7.19	2.89	0.29	40.9	3.71	0.000
	2	15.89	10.90	4.99	0.31	85.2	3.08	0.001
	10	38.56	29.39	9.17	0.24	450.7	1.06	0.144
Munich RE	1	7.49	6.06	1.42	0.19	17.9	4.14	0.000
	2	10.92	8.75	2.17	0.20	35.0	3.22	0.001
	10	21.17	19.92	1.25	0.06	124.4	0.52	0.300
Siemens	1	8.89	6.85	2.04	0.23	28.5	3.75	0.000
	2	12.07	9.32	2.75	0.23	39.0	3.70	0.000
	10	31.33	26.09	5.24	0.17	95.7	2.87	0.002

TABLE 5

Mean scores for conditional and unconditional VaR estimation ($\alpha = 0.01$) for the log-returns of several stocks (date values starting from at least 2001-01-02 resulting in a value of $N \geq 3211$ in each row)

both Bernoulli-sequences with success probabilities α.

Simulation

We simulate the series $(\mathbf{R}_t)$ from a bivariate DCC-GARCH-model of Engle (2002), where

$$
\begin{aligned}
\mathbf{R}_t &= H_t^{1/2}\boldsymbol{\epsilon}_t, \qquad \text{with} \quad \boldsymbol{\epsilon}_t \text{ i.i.d.} \sim N(\mathbf{0}, I_2),\\
H_t &= D_t C_t D_t, \qquad D_t = \operatorname{diag}(\sigma_{t,1}, \sigma_{t,2}),\\
\sigma_{t,i}^2 &= \kappa_i + \phi_i^2 R_{t-1,i}^2 + \beta_i \sigma_{t,i-1}^2,\\
C_t &= \operatorname{diag}(q_{t;1,1}^{-1/2}, q_{t;2,2}^{-1/2})\, Q_t \operatorname{diag}(q_{t;1,1}^{-1/2}, q_{t;2,2}^{-1/2}), \qquad Q_t = \big(q_{t;j,k}\big)_{j,k=1,2},\\
Q_t &= (1-\gamma-\eta)\bar{Q} + \gamma \boldsymbol{u}_{t-1}\boldsymbol{u}_{t-1}^T + \eta Q_{t-1},\\
\boldsymbol{u}_t &= \big(R_{t,1}/\sigma_{t,1}, R_{t,2}/\sigma_{t,2}\big)^T, \qquad \bar{Q} = \operatorname{cov}(\boldsymbol{u}_t),
\end{aligned}
\tag{18}
$$

and the parameters are chosen according to the scenarios listed in Table 6, $\mathbf{w} = (1/2, 1/2)^T$ and $\alpha = 0.01$.

a. Again, we first approximate the true mean score of the (approximate) ideal forecasts by sample averages $\widehat{m}_{N,\mathcal{G}}$ and $\widehat{m}_{N,\mathcal{F}}$ based on a single huge sample of size $N = 500{,}000$. In the multivariate case for $\hat{Y}_{t+1,\mathcal{G}}$ and $\widehat{m}_{N,\mathcal{G}}$, we use the true parameters of the DCC-GARCH-model and the exact forecast distribution $N(0, \mathbf{w}^T H_t \mathbf{w})$. For $\hat{Y}_{t+1,\mathcal{F}}$ and $\hat{\mu}_{N,\mathcal{F}}$, we first determine an appropriate model for the series of (Y_t) within the class of GARCH(p,q)-models, and then use one-step forecasts within this univariate GARCH-model. Even though the class of multivariate GARCH models is not closed under aggregation, it turns out that a simple GARCH$(1,1)$-model with normal innovations works surprisingly well. The results can be found in Table 7. While in all scenarios, the difference between the average scores is significant due to the high sample sizes, the relative reduction in mean score is small with maximal values of 0.06.

Simulations for a class of regime-switching models which are closed under aggregation led to similar results.

Config	κ_1	κ_2	ϕ_1	ϕ_2	β_1	β_2	$\bar{q}_{21}$	γ	η
1	0.0030	0.0010	0.400	0.050	0.590	0.930	0.10	0.01	0.98
2	0.0025	0.0015	0.390	0.060	0.600	0.920	0.30	0.02	0.97
3	0.0100	0.0070	0.200	0.180	0.790	0.800	0.30	0.08	0.91
4	0.0200	0.0010	0.100	0.300	0.890	0.680	0.35	0.10	0.89
5	0.0030	0.0010	0.400	0.005	0.590	0.975	0.60	0.01	0.98
6	0.0090	0.0080	0.200	0.010	0.790	0.970	0.75	0.05	0.94
7	0.0028	0.0031	0.300	0.500	0.690	0.480	0.88	0.01	0.98

TABLE 6

Configurations for the simulation of the DCC-GARCH-model ($N = 500{,}000$, $\alpha = 0.01$, $w_1 = 0.5$, $w_2 = 0.5$)

Config	Mean Scores		Diff. ($= M_N$)	Rel. diff.			
	$\hat{m}_{N,\mathcal{F}}$	$\hat{m}_{N,\mathcal{G}}$	$\hat{m}_{N,\mathcal{F}} - \hat{m}_{N,\mathcal{G}}$	$M_N/\hat{m}_{N,\mathcal{F}}$	$\hat{\sigma}$	T_N	$\Pr(> T_N)$
1	0.527	0.495	0.031	0.06	1.1	20.93	< 0.001
2	0.580	0.556	0.024	0.04	0.9	19.02	< 0.001
3	1.330	1.322	0.007	0.01	0.6	9.18	< 0.001
4	1.727	1.725	0.002	0.00	0.3	4.96	< 0.001
5	0.595	0.574	0.021	0.04	1.2	12.73	< 0.001
6	1.648	1.628	0.020	0.01	1.3	11.12	< 0.001
7	0.666	0.662	0.003	0.00	0.3	6.33	< 0.001

TABLE 7

Mean scores for univariate and multivariate VaR estimation ($N = 500{,}000$, $\alpha = 0.01$); for parameter configurations 1 to 7, cf. Table 6

	Config						
N	1	2	3	4	5	6	7
250	0.099	0.086	0.091	0.094	0.080	0.088	0.044
500	0.112	0.084	0.073	0.051	0.075	0.083	0.055
1000	0.159	0.110	0.037	0.044	0.111	0.092	0.058
1500	0.232	0.183	0.059	0.045	0.172	0.129	0.088
2000	0.305	0.219	0.076	0.048	0.201	0.140	0.085
4000	0.567	0.418	0.100	0.038	0.403	0.245	0.104
6000	0.707	0.528	0.116	0.035	0.545	0.322	0.104

TABLE 8

Power of the test (at the 0.05 level) for univariate and multivariate VaR estimation ($\alpha = 0.01$); for parameter configurations, cf. Table 6

b. Next, we investigate the power of the resulting DM test for realistic sample sizes when taking into account estimation effects. Again, we base estimation on a rolling window of sizes $R_{\text{wind}} = 500$, and proceed as in part b. above. The resulting power estimates for test levels 0.05 and 0.1, which are reasonably high at least for higher sample sizes, can be found in Tables 8 and 9.

Application

We proceed with an application to portfolios consisting of two stocks. Let $S_{t,i}$, $i = 1, 2$, denote

	Config						
N	1	2	3	4	5	6	7
250	0.195	0.177	0.162	0.143	0.169	0.159	0.106
500	0.262	0.208	0.143	0.090	0.181	0.166	0.131
1000	0.327	0.247	0.104	0.091	0.251	0.196	0.139
1500	0.419	0.341	0.140	0.107	0.308	0.234	0.161
2000	0.496	0.382	0.158	0.103	0.358	0.266	0.155
4000	0.737	0.589	0.181	0.096	0.554	0.399	0.168
6000	0.835	0.706	0.206	0.093	0.710	0.468	0.184

TABLE 9

Power of the test (at the 0.1 *level) for univariate and multivariate VaR estimation* ($\alpha = 0.01$)*; for parameter configurations, cf. Table 6*

the price of stock i at time t (daily closure). Consider the relative returns

$$R_{t,i} = \frac{S_{t,i} - S_{t-1,i}}{S_{t-1,i}}, \qquad i = 1, 2.$$

Let $\lambda_{t,i}$ denote the amount held from stock i from time t to time $t+1$, and let $V_t = \lambda_{t,1}\, S_{t,1} + \lambda_{t,2}\, S_{t,2}$. Then for the portfolio return (Y_t),

$$Y_{t+1} = \sum_{i=1}^{2} R_{t+1,i}\, \lambda_{t,i}\, \frac{S_{t,i}}{V_t},$$

so that in order to obtain the constant weights w_i, $i = 1, 2$, on the basis of returns, we choose $\lambda_{t,i} = w_i V_t / S_{t,i}$ with initial value $V_0 = 1$. We model the series $(\mathbf{R}_t)$, $\mathbf{R}_t = (R_{t,1}, R_{t,2})^T$, by a DCC-GARCH-model as specified above and the univariate series (Y_t) of portfolio returns by a simple GARCH$(1,1)$-model. In both cases, at time t using a rolling window we base the estimation on the last $R_{\text{wind}} = 500$ observations.
The results are contained in Table 11. The difference in estimated mean scores, which is negative in about $5/12$ cases under consideration, is not significantly $\neq 0$ each time.

Abbreviation	Full name	Abbreviation	Full name
ADS.DE	Adidas	FRE.DE	Fresenius VZ
ALV.DE	Allianz	HEI.DE	Heidelbergcement
BEI.DE	Beiersdorf	HEN3.DE	Henkel VZ
BMW.DE	BMW	MRK.DE	Merck
DAI.DE	Daimler	MUV2.DE	Munich RE
DBK.DE	Deutsche Bank	RWE.DE	RWE
EOAN.DE	E.ON	SIE.DE	Siemens
FME.DE	Fresenius Medical Care		

TABLE 10

List of share name abbreviations used in Table 11

Conclusions

Using the models under consideration, there seems to be small potential for improvement by using the multivariate DCC-model for the individual risk factors instead of the simple GARCH$(1,1)$-model for the portfolio returns. However, further investigations with distinct multivariate time series models would be required.

Share N° 1 Abbr.	Share N° 2 Abbr.	Corr.	Mean Scores $\hat{m}_{N,\mathcal{F}}$	$\hat{m}_{N,\mathcal{G}}$	Diff. ($= M_N$) $\hat{m}_{N,\mathcal{F}} - \hat{m}_{N,\mathcal{G}}$	Rel. diff. $M_N/\hat{m}_{N,\mathcal{F}}$	$\hat{\sigma}$	T_N	Pr($> T_N$)
FME.DE	HEI.DE	0.104	4.80	4.76	0.04	0.01	7.1	0.30	0.383
ADS.DE	FME.DE	0.186	3.98	4.00	−0.02	−0.01	3.3	−0.34	0.632
ADS.DE	BEI.DE	0.204	4.02	4.02	0.00	0.00	3.1	0.05	0.480
FME.DE	MRK.DE	0.218	4.51	4.50	0.00	0.00	3.7	0.02	0.492
FRE.DE	HEN3.DE	0.227	3.91	3.84	0.07	0.02	3.2	1.04	0.150
FME.DE	HEN3.DE	0.266	3.84	3.87	−0.03	−0.01	2.6	−0.51	0.694
DAI.DE	SIE.DE	0.654	5.78	5.90	−0.12	−0.02	3.5	−1.76	0.961
EOAN.DE	RWE.DE	0.680	6.24	6.25	−0.01	−0.00	6.7	−0.04	0.517
BMW.DE	DAI.DE	0.719	5.54	5.56	−0.02	−0.00	2.4	−0.40	0.654
ALV.DE	DBK.DE	0.744	6.01	6.00	0.01	0.00	1.8	0.25	0.401
ALV.DE	MUV2.DE	0.766	5.24	5.22	0.03	0.00	1.8	0.72	0.237

TABLE 11

Mean scores for univariate and multivariate VaR estimation ($\alpha = 0.01$) for the log-returns of several stocks (date values starting from at least 2003-01-01); for full names of shares cf. Table 10

5. Concluding remarks. Additional information should lead to better forecasts, at least if the forecasting mechanism is ideal, that is based on the true conditional distribution. But how can the improvement of an increase in information on the forecast, e.g. the mean, a quantile or the whole predictive distribution, be quantified, what exactly is improved?
The answer that we give in this paper is in terms of the expected loss (score) under a strictly consistent scoring function or rule, which is attuned to the predicted parameter. This interpretation is particularly attractive if the expected loss is by itself of interest. For instance, for the value at risk (a quantile), we show that the expected loss under an appropriate scoring function turns out to be the expected shortfall.
While for ideal forecasts, additional information is thus always useful or at least not harmful, this is apparently no longer true if information, e.g. data, needs to be processed by a statistician in terms of model building, selection and estimation before making predictions. For example, in our application on value at risk prediction for log-returns, it turned out that a multivariate modeling of individual risk factors often performs worse than a simple univariate modeling of the portfolio returns.
Thus, the development of model selection criteria with the aim of optimal prediction of a certain parameter under a specific scoring function, such as the AIC for the mean and squared error in regression models, should be a major issue of future research.

Acknowledgements. The authors would like to thank the editor Tilmann Gneiting as well as three anonymous referees for helpful comments and for pointing out several relevant references, and Steffen Dereich for helpful discussions. Hajo Holzmann gratefully acknowledges financial support from the DFG, grant Ho 3260/3-1.

References.

ACERBI, C. and TASCHE, D. (2002). On the coherence of expected shortfall. *J. Banking Finance* **26** 1487–1503.

BAO, Y., LEE, T.-H. and SALTOĞLU, B. (2006). Evaluating predictive performance of value-at-risk models in emerging markets: a reality check. *J. Forecast.* **25** 101–128. MR2226780

BERKOWITZ, J., CHRISTOFFERSEN, P. F. and PELLETIER, D. (2011). Evaluating value-at-risk models with desk-level data. *Management Science* **57** 2213–2227.

BRÖCKER, J. (2009). Reliability, sufficiency, and the decomposition of proper scores. *Q. J. Roy. Meteor. Soc.* **135** 1512–1519.

CHRISTOFFERSEN, P. F. (1998). Evaluating interval forecasts. *Internat. Econom. Rev.* **39** 841–862.

CHRISTOFFERSEN, P. F. (2009). Value-at-Risk Models. In *Handbook of Financial Time Series* (T. Mikosch, J. P. Kreiß, R. A. Davis and T. G. Andersen, eds.) 753–766. Springer, Berlin Heidelberg. http://dx.doi.org/10.1007/978-3-540-71297-8_33.

DEGROOT, M. H. and FIENBERG, S. E. (1983). The Comparison and Evaluation of Forecasters. *J. Roy. Stat. Soc. Ser. D (The Statistician)* **32** 12–22.

DIEBOLD, F. X. (2012). Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of Diebold-Mariano tests. *NBER Working Paper No. 18391*.

DIEBOLD, F. X. and MARIANO, R. S. (1995). Comparing predictive accuracy. *J. Business Economic Statist.* **13** 253–263.

DURRETT, R. (2005). *Probability: Theory and Examples*, 3rd ed. *Duxbury Avanced Series*. Thomson Brooks/Cole, Belmont, CA.

ENGLE, R. (2002). Dynamic conditional correlation: a simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *J. Bus. Econom. Statist.* **20** 339–350. MR1939905

ESCANCIANO, J. C. and OLMO, J. (2011). Robust backtesting tests for value-at-risk models. *J. Financ. Economet.* **9** 132–161.

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

GNEITING, T. (2011). Making and evaluating point forecasts. *J. Amer. Statist. Assoc.* **106** 746–762. MR2847988 (2012j:62355)

GNEITING, T., BALABDAOUI, F. and RAFTERY, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **69** 243–268. MR2325275

GNEITING, T. and RAFTERY, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.* **102** 359–378. MR2345548

GNEITING, T. and RANJAN, R. (2011). Comparing density forecasts using threshold- and quantile-weighted scoring rules. *J. Bus. Econom. Statist.* **29** 411–422. MR2848512 (2012j:62356)

HEINRICH, C. (2013). The mode functional is not elicitable. *Biometrika*. Advance Access published November 28, 2013. http://dx.doi.org/10.1093/biomet/ast048.

JORION, P. (2006). *Value-at-Risk: The New Benchmark for Managing Financial Risk*. Mc-Graw Hill, New York.

KLENKE, A. (2008). *Probability theory. Universitext*. Springer-Verlag London Ltd., London. A comprehensive course, Translated from the 2006 German original. MR2372119 (2009f:60001)

MCNEIL, A. J., FREY, R. and EMBRECHTS, P. (2005). *Quantitative Risk Management. Princeton Series in Finance*. Princeton University Press, Princeton, NJ. Concepts, Techniques and Tools. MR2175089 (2006d:91005)

MITCHELL, J. and WALLIS, K. F. (2011). Evaluating density forecasts: forecast combinations, model mixtures, calibration and sharpness. *J. Appl. Econometrics* **26** 1023–1040. MR2843116

NEWEY, W. K. and WEST, K. D. (1987). A simple, positive semidefinite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* **55** 703–708. MR890864 (88g:62111)

PATTON, A. J. and TIMMERMANN, A. (2012). Forecast rationality tests based on multi-horizon bounds. *J. Bus. Econom. Statist.* **30** 1–17. MR2899176

ROCKAFELLAR, R. T. and URYASEV, S. (2000). Optimization of conditional value-at-risk. *J. Risk* **2** 21–41.

TSYPLAKOV, A. (2011). Evaluating density forecasts: a comment. *MPRA Paper No. 31233*. http://mpra.ub.uni-muenchen.de/31233.

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes. Springer Series in Statistics*. Springer-Verlag, New York. With Applications to Statistics. MR1385671 (97g:60035)

APPENDIX

Proofs. We start with the following well-known fact, which we prove for lack of reference.

LEMMA 9. *Let $(\Omega, \mathcal{F}, P)$ be a probability space, and let $G_{\mathcal{F}} : \Omega \times \mathcal{B} \to [0,1]$ be a Markov kernel for which $G_{\mathcal{F}}(\omega, \cdot) \in \Theta$ for all $\omega \in \Omega$. Then $G_{\mathcal{F}} : \Omega \to \Theta$, $\omega \mapsto G_{\mathcal{F}}(\omega, \cdot)$ is $\mathcal{F} - \mathcal{B}(\Theta)$ measurable.*

PROOF. For a fixed continuous, bounded function $f : D \to \mathbb{R}$, the map

$$\omega \mapsto \int_{\mathbb{R}} f(x)\, G_{\mathcal{F}}(\omega; dx) \tag{19}$$

is $\mathcal{F}-\mathcal{B}$-measurable, see Klenke (2008, Theorem 8.37). The weak topology on Θ may be metrized by

$$d(\mu,\nu)=\sup_n\left\{\Big|\int f_n d\mu-\int f_n d\nu\Big|:\mu,\nu\in\Theta\right\},$$

where (f_n) is a appropriate sequence of bounded, continuous functions on D, see van der Vaart and Wellner (1996, Theorem 1.12.2).
The metric space (Θ,d) is separable, see Klenke (2008, p. 252). Therefore, for the measurability of $G_{\mathcal{F}}$, it suffices to show that the preimage of every closed ball $B_\epsilon(\mu)$, $\epsilon>0$, $\mu\in\Theta$ in the metric d, under $G_{\mathcal{F}}$ is in $\mathcal{F}$. Now,

$$B_{n,\epsilon}(\mu)=\left\{\omega\in\Omega\colon\Big|\int f_n(x)G_{\mathcal{F}}(\omega;dx)-\int f_n(x)d\mu(x)\Big|\le\epsilon\right\}\in\mathcal{F}$$

by (19), and hence also

$$G_{\mathcal{F}}^{-1}\big(B_\epsilon(\mu)\big)=\bigcap_n B_{n,\epsilon}(\mu)\in\mathcal{F}.$$

□

Proof of Theorem 1. Let $\mu_{Y|\mathcal{F}}$ denote the conditional distribution with corresponding conditional distribution functions $F_{Y|\mathcal{F}}$. Since by the above lemma, the map $\mu_{Y|\mathcal{F}}:\Omega\to\Theta$, $\omega\mapsto\mu_{Y|\mathcal{F}}(\omega,\cdot)$, is $\mathcal{F}-\mathcal{B}(\Theta)$-measurable, and since by assumption T is $\mathcal{B}(\Theta)-\mathcal{B}$-measurable, it follows that $\hat{Y}(\omega)=T\circ\mu_{Y|\mathcal{F}}(\omega;\cdot)$ is an $\mathcal{F}$-measurable random variable.
For $P-a.e.\ \omega\in\Omega$,

$$E\big(S(Z,Y)|\mathcal{F}\big)(\omega)=\int_{\mathbb{R}}S\big(Z(\omega),y\big)\,F_{Y|\mathcal{F}}(\omega,dy).$$

Since S is strictly consistent, for all $\omega\in\Omega$ we have

$$\int_{\mathbb{R}}S\big(\hat{Y}(\omega),y\big)\,F_{Y|\mathcal{F}}(\omega,dy)\le\int_{\mathbb{R}}S\big(Z(\omega),y\big)\,F_{Y|\mathcal{F}}(\omega,dy)$$

with equality if and only if $Z(\omega)=\hat{Y}(\omega)$. The second statement follows by taking expectated values. □

Proof of Corollary 2. The proof of the first statement of (5) is immediate from Theorem 1, since $\hat{Y}_{\mathcal{F}}$ is also $\mathcal{G}$-measurable. For the second, take conditional expectation w.r.t. $\mathcal{F}$. Since for a non-negative random variable Z, $Z=0$ a.s. if and only if $E(Z|\mathcal{F})=0$ a.s., the second conclusion follows. For the third, take unconditional expectation. □

Proof of Theorem 3. Set $X(\omega)=\boldsymbol{S}\big(G_{\mathcal{F}}(\omega,\cdot),Y(\omega)\big)$. By Lemma 9, X is measurable. Then for $P-a.e.\ \omega\in\Omega$,

$$E\big(X|\mathcal{F}\big)(\omega)=\int_{\mathbb{R}}\boldsymbol{S}\big(G_{\mathcal{F}}(\omega,\cdot),y\big)\,F_{Y|\mathcal{F}}(\omega,dy).$$

Since $\boldsymbol{S}$ is strictly proper, for all $\omega\in\Omega$ we have

$$\int_{\mathbb{R}}\boldsymbol{S}\big(F_{Y|\mathcal{F}}(\omega,\cdot),y\big)\,F_{Y|\mathcal{F}}(\omega,dy)\le\int_{\mathbb{R}}\boldsymbol{S}\big(G_{\mathcal{F}}(\omega,\cdot),y\big)\,F_{Y|\mathcal{F}}(\omega,dy)$$

with equality if and only if the distributions $F_{Y|\mathcal{F}}(\omega,\cdot)$ and $G_{\mathcal{F}}(\omega,\cdot)$ coincide. This proves the first part of the theorem, the second follows by taking unconditional expected values. The final statement is a standard fact of probability. □

Proof of Theorem 5. Set

$$W_n = \frac{1}{n}\sum_{k=1}^{n}\Big(S(\hat{Y}^{(h)}_{k,\mathcal{F}}, Y_k) - S(\hat{Y}^{(h)}_{k,\mathcal{G}}, Y_k)\Big) = \frac{1}{n}\sum_{k=1}^{n} Z_k.$$

Under an alternative, Corollary 2, (5), 3$^{\text{rd}}$ statement, implies that $EZ_1 > 0$, and the ergodic theorem then implies $\sqrt{n}W_n \to \infty$ P–a.s.
From (13) with $O_P(\sqrt{n})$, $\sqrt{n}\,(M_n - W_n) = O_P(1)$, therefore, $\sqrt{n}M_n \to \infty$ $P - a.s.$ as well.
Under the null hypothesis, from Corollary 2, (5), 1$^{\text{st}}$ statement with equality implies that $E(Z_n|\mathcal{G}_{n-h}) = 0$ for all n. Therefore, setting $\|X\|_2 = (EX^2)^{1/2}$, we have that

$$\sum_{n=0}^{\infty}\|E(Z_0|\mathcal{G}_{-n})\|_2 = \sum_{n=0}^{h-1}\|E(Z_0|\mathcal{G}_{-n})\|_2 < \infty,$$

and from the CLT for stationary sequences, see Durrett (2005, Theorem (7.6), p. 416).

$$\sqrt{n}W_n \xrightarrow{d} N\big(0, \sigma^2\big),$$

where σ^2 is as in (14). From (13) with $o_P(\sqrt{n})$, $\sqrt{n}\,(M_n - W_n) = o_P(1)$, therefore, asymptotic normality holds true for $\sqrt{n}M_n$ as well.

□

Proof of Proposition 6. We only show the implication 1. ⇒ 2. By independence, we have that for $P - a.e.\ \omega \in \Omega$,

$$F_{Y|\mathcal{F}}\big(\omega, (Z(\omega), \infty)\big) = P(I = 1|\mathcal{F})(\omega) = P(I = 1) = 1 - \alpha,$$

so that $Z(\omega) = q_\alpha\big(F_{Y|\mathcal{F}}(\omega, \cdot)\big)$. □

Proof of Corollary 7. This follows from the fact that the (I_n) are independent if and only if for all n, I_{n+1} and $\sigma(I_k; k \le n)$ are independent. □

Proof of Proposition 8. For a strictly increasing, continuous distribution function F, a simple calculation gives that

$$E_F\big(S^*(q_\alpha(F), Y)\big) = -\frac{1}{\alpha}\int_{-\infty}^{q_\alpha(F)} y\, dF(y).$$

Therefore, (17) follows from (5). □

E-MAIL: holzmann@mathematik.uni-marburg.de E-MAIL: eulert@mathematik.uni-marburg.de