Evaluating Density Forecasts with an Application to Stock Market Returns

Gabriela de Raaij
(Oesterreichische Nationalbank)

Burkhard Raunig
(Oesterreichische Nationalbank)

Discussion paper 08/02

Economic Research Centre

of the Deutsche Bundesbank

February 2002

Deutsche Bundesbank, Wilhelm-Epstein-Strasse 14, 60431 Frankfurt am Main,
Postfach 10 06 02, 60006 Frankfurt am Main

Tel +49 69 95 66-1
Telex within Germany 4 1 227, telex from abroad 4 14 431, fax +49 69 5 60 10 71

Please address all orders in writing to: Deutsche Bundesbank,
Press and Public Relations Division, at the above address or via fax No. +49 69 95 66-30 77

ISBN 3–935821–05–0

Abstract

Density forecasts have become quite important in economics and finance. For example, such forecasts play a central role in modern financial risk management techniques like Value at Risk. This paper suggests a regression based density forecast evaluation framework as a simple alternative to other approaches. In simulation experiments and an empirical application to in- and out-of-sample one-step-ahead density forecasts of daily returns on the S&P 500, DAX and ATX stock market indices, the regression based evaluation strategy is compared with a recently proposed methodology based on likelihood ratio tests. It is demonstrated that misspecifications of forecasting models can be detected within the proposed regression framework. It is further demonstrated that the likelihood ratio methodology without additional misspecification tests has no power in many practical situations and therefore frequently selects incorrect forecasting models. The empirical results provide some evidence that GARCH-t models provide good density forecasts. The results further suggest that extensions of statistical models with fat-tailed conditional distributions to models that incorporate higher order conditional moments beyond the conditional variance might be appropriate to capture the empirical regularities in financial time series in some cases.

Key words: Density forecasting, Forecast evaluation, Risk management, GARCH-models

JEL Classification: G10, C52, C53

Zusammenfassung

Zur Beurteilung der Voraussage von Dichten und einer Anwendung auf Aktienmärkte.

Die Voraussagen von Dichten ist in verschiedenen ökonomischen Fragestellungen sehr wichtig geworden. Solche Voraussagen spielen zum Beispiel eine wichtige Rolle bei modernen Methoden des Risikomanagements im Finanzsektor. Dieses Papier schlägt vor, Dichte-Prognosen mithilfe einer Methode zu beurteilen, die auf einem Regressionsansatz beruht. In Simulationsexperimenten und empirischen Anwendungen auf Dichte-Prognosen für tägliche Erträge verschiedener Aktienindices (S&P 500, DAX, ATX) wird diese Methode mit einer verglichen, die auf likelihood ratio Tests beruht und die erst neulich vorgeschlagen wurde. Es zeigt sich, dass Fehlspezifikationen der Prognosemodelle mithilfe der hier vorgeschlagenen Methode entdeckt werden können. Dagegen hat die Methode, die auf likelihood ratio Test beruht, ohne zusätzliche Tests auf Fehlspezifikation in vielen praktischen Fällen keine Macht. Die empirischen Ergebnisse deuten darauf hin, dass GARCH-t-Modelle gute Dichte-Prognosen liefern. Weiterhin wird gezeigt, dass Erweiterungen von statistischen Modellen mit Verteilungen mit dicken Enden zu Modellen, die höhere Momente einbeziehen, geeignet sein können, um in manchen Fällen empirische Regelmäßigkeiten in Finanzzeitreihen abzubilden.

Table of Contents

List of Tables and Figures

Tables

Evaluating Density Forecasts with an Application to Stock Market Returns *

1 Introduction

A density forecast is a forecast of the entire probability distribution of a random variable. Recently, such forecasts have become quite important in the financial industry because they form the backbone of modern risk measures like Value at Risk (VaR) which are derived from forecasts of entire profit/loss distributions of financial portfolios (for details on VaR, see Jorion, 1997). Apart from risk management, density forecasts have also come to play a role in macroeconomic forecasting. Density forecasts of inflation are assessed in Diebold, Tay and Wallis (1999), density forecasts of output growth and unemployment are examined in Clements and Smith (2000) and Kaufmann (2000) evaluates the statistical adequacy of a dynamic Markov switching factor model for the business cycle using the predictive densities implied by the model (for a survey about density forecasting, see Tay and Wallis, 2000). Given the rapidly growing importance of density forecasts for economic forecasting in general and risk management in particular, techniques to evaluate the quality of density forecasts are of vital practical importance.

Recently, Crnkovic and Drachman (1997) and Diebold, Gunther and Tay (1998) have introduced methodologies to evaluate the accuracy of density forecasts based on a probability integral transformation in Rosenblatt (1952). Applied to the realizations of a stochastic process, the transformation implies iid U(0,1) data if a sequence of density forecasts coincides with the sequence of true conditional densities. Frühwirth-Schnatter (1996) and Berkowitz (2000) extend this framework by utilizing a second transformation that implies iid N(0,1) data if a sequence of density forecasts is correct. Whereas Frühwith-Schnatter proposes certain indices to explore the adequacy of forecasting models, Berkowitz suggests a likelihood ratio (LR) framework to test for iid N(0,1). He further finds that the LR-framework is quite powerful even in small samples. However, the LR-framework maintains the assumption of normality and therefore does not cover the complete hypothesis.

---

* This working paper was previously presented at the regular joint research workshop of the Deutsche Bundesbank and the Oesterreichische Nationalbank. It is also available as Oesterreichische Nationalbank working paper No. 59. The opinions expressed do not necessarily reflect those of the Oesterreichische Nationalbank or the Deutsche Bundesbank.

Unfortunately, under standard statistical assumptions about a forecasting model situations may arise where deficiencies in density forecasts cannot be detected within the LR-framework. An important case are density forecasts derived from GARCH-models with correctly specified first- and second moments estimated with quasi-maximum likelihood methods. This paper demonstrates that deficient density forecasts derived from such models may not be detected within the LR-framework. It is further shown that the LR-methodology also has no power under the weaker condition of a correct specification of the conditional mean of a forecasting model, if normally distributed density forecasts are assumed.

To overcome these problems, this paper proposes a regression framework in conjunction with tests for normality to evaluate the quality of density forecasts. The approach is motivated through a probabilistic reduction argument (Spanos, 1999, ch.15) and covers the alternative hypotheses of the LR-tests proposed in Berkowitz (2000) as a special case. Given a reasonable sample size, the regression framework does not require the assumptions of normality and homoskedasticity in tests concerning the correlation structure of a transformed series and the additional tests for normality provide further important information about deficiencies of density forecasts and hence about misspecifications of the models that were used to generate the forecasts. Since neglected conditional volatility dynamics in a forecasting model induces heteroskedasticity into the transformed series used for density forecast evaluation, tests that help to identify such effects are an integral part of the evaluation framework. Simulation experiments indicate that the proposed methodology has good statistical properties.

In an empirical application the regression methodology is used to evaluate in-sample and out-of-sample one-step-ahead density forecasts from econometric models that are popular in the financial industry and the results are compared with the results from the LR-approach. The different forecasting models are applied to daily stock market returns from the S&P 500, the DAX and the ATX. The empirical results provide some support for GARCH-models with fat tailed distributed errors for the purpose of density forecasting and GARCH-models in general for the purpose of volatility forecasting. The results further suggest that for financial return series an adequate model for the relevant conditional moments as well as proper distributional assumptions are needed to produce good density forecasts.

The rest of the paper is organized as follows. Section 2 covers some theory about density forecast evaluation, reviews the LR-framework, discusses properties of transformed series obtained from misspecified forecasting models and provides conditions under which LR-tests will fail to detect incorrect density forecasts. Section 3 outlines the regression based

evaluation approach. Section 4 reports simulation experiments concerning the size and the power of LR-tests and the regression based evaluation methodology. The data and the models used in the empirical study are presented in section 5. Section 6 describes the setting of the forecasting experiments and discusses the in- and out-of-sample forecast evaluation results. Section 7 comprises the concluding remarks. Proofs are collected in an appendix.

2 Density Forecast Evaluation

Let $\{x_t\}_{t = 1,..., m}$ be a time series generated from the series of conditional densities $\{f(x_t| I_{t-1})\}_{t = 1,..., m}$ where I_{t-1} denotes the information set available at time t-1 and let $\{p(x_t| I_{t-1})\}_{t = 1,..., m}$ be a series of one-step-ahead density forecasts for $\{x_t\}_{t = 1,..., m}$ (in what follows, $f_t(x_t)$ and $p_t(x_t)$ are sometimes used as shorthand notations for the true and the predicted conditional densities, respectively). Assume that a series of one-step-ahead density forecasts has been generated. Such forecasts can be evaluated through a probability integral transformation (Rosenblatt, 1952) applied to each observed x_t with respect to its predicted density $p_t(x_t)$. The probability integral transformation for a single x_t is given by

$$(1) \qquad z_t = \int_{-\infty}^{x_t} p_t(u)du = P_t(x_t).$$

Diebold, Gunther and Tay (1998) show that a transformed series $\{z_t\}_{t = 1,...,m}$ is iid U(0,1) if a series of one-step-ahead density forecasts $\{p_t(x_t)\}_{t = 1, ..., m}$ coincides with the series of the true densities $\{f_t(x_t)\}_{t = 1,..., m}$. This result can be further exploited to evaluate multivariate density forecasts- and multi-step ahead forecasts, respectively (Diebold, Hahn and Tay, 1999, Clements and Smith, 2000). It is also worth noting that this result does in no way depend on how the density forecasts were generated. Correct density forecasts, however obtained, imply a transformed series that is iid U(0,1).

Diebold et al. suggest graphical methods to assess the iid U(0,1) property of transformed data and Crnkovic and Drachman (1997) advocate Kupier's statistics to test for uniformity and Brock-Dechert-Scheinkman (BDS) tests for iid. Berkowitz (2000) emphasizes that nonparametric tests require rather large sample sizes to be reliable. He therefore suggests a further well known transformation (the so called quantile transformation) to the individual z_t's. The transformation for a single z_t is given by

$$(2) \qquad n_t = F_N^{-1}(z_t).$$

This transformation produces data that are standard normal if z_t is U(0,1) and F_N^{-1} is the inverse of a standard normal distribution function. If a series of z_t's is iid U(0,1) it follows from the iid property of the z-series that the corresponding n-series must also be iid N(0,1). Berkowitz proposes likelihood-ratio tests against the first order autoregressive alternative

(3) $$n_t - \mu = \rho(n_{t-1} - \mu) + \varepsilon_t,$$

to test for iid N(0,1) data. In this framework a test for independence is given by (3a) and a joint test for independence, a mean of zero and a variance of one is given by (3b)

(3a) $$LR_1 = -2\left(L(\hat{\mu}, \hat{\sigma}^2, 0) - L(\hat{\mu}, \hat{\sigma}^2, \hat{\rho})\right) \sim \chi^2(1)$$

(3b) $$LR_2 = -2\left(L(0,1,0) - L(\hat{\mu}, \hat{\sigma}^2, \hat{\rho})\right) \sim \chi^2(2),$$

where σ^2 is the variance of ε_t and L(.) denotes a Gaussian log-likelihood function. A test concerning $\mu = 0$ can be constructed analogously to (3a). In simulation experiments he demonstrates that the test statistics have good small sample properties. However, Berkowitz himself points out that the LR tests have only power to detect non normality through the first two moments of the distribution and that additional distributional tests might be useful.

There are indeed good reasons to examine an n-series of a forecasting model for normality and also for heteroskedasticity because the LR-tests outlined above maintain the assumption of normality and do not cover the possibility of heteroskedasticity. If density forecasts are deficient in such a way that they do not lead to a violation of $\mu = 0$, $\rho = 0$ and $\sigma = 1$, then these deficiencies will not be detected within the LR-framework and an incorrect forecasting model may be selected. Propositions 1 and 2 below state that this will happen under standard statistical assumptions about a forecasting model.

Proposition1: Assume that the forecasting model can be represented in the form $X_t = \mu(I_{t-1}) + \sigma(I_{t-1})Y_t$, where $\mu(I_{t-1})$ is the conditional mean and $\sigma(I_{t-1})$ is the conditional standard deviation of X_t. The random variable Y_t is iid with some arbitrary distribution D(0,1) with zero mean and unit variance. Further assume that the forecasted densities $p(X_t \mid I_{t-1})$ adequately capture the first two conditional moments of X_t. Then the n-series implied by the forecasting model will also be iid D(0,1) if the forecasted densities $\{p(X_t \mid I_{t-1})\}_{t = 1, \dots, m}$ are assumed to be normal densities.

Proof: see appendix

Proposition 2: Assume a forecasting model of the form $X_t = \mu(I_{t-1}) + \sigma(I_{t-1})Y_t$, where $\mu(I_{t-1})$ is the conditional mean and $\sigma(I_{t-1})$ is the conditional standard deviation of X_t. Further assume that the (constant) unconditional standard deviation σ exists. Then the n-series $\{n_t\}_{t = 1, \dots, m}$ resulting from the forecasting model is

a) conditionally heteroskedastic

b) uncorrelated, has conditional mean and unconditional mean 0

c) and has unconditional standard deviation 1

if the density forecasts $\{p(X_t \mid I_{t-1})\}_{t = 1, \dots, m}$ adequately capture the conditional mean, are assumed to be normal and are based on unconditional standard deviation σ.

Proof: see appendix

Propositions 1 implies that the LR-tests given in (3a) and (3b) and other tests that do not cover the distributional part of the iid N(0,1) hypotheses of correct density forecasts will have no power in detecting incorrect density forecasts if the stated conditions apply. An important practical case arises in the context of quasi-maximum likelihood estimation (QML) of GARCH models. It is well known that under mild regularity conditions the parameters of a GARCH model estimated under the incorrect assumption of a normal distribution are consistent if the conditional mean- and the conditional variance functions are correctly specified (for details, see Bollerslev and Wooldridge, 1992, and Lumsdaine, 1996). Hence a GARCH model might approximate the first two conditional moments quite well, but deliver poor density forecasts under the incorrect assumption of normality. Proposition 1 states that in such a situation the LR-tests will virtually never reject the null-hypotheses of correct density forecasts because the iid, mean 0 and variance 1 property of the derived n-series will not be violated. Without additional tests for normality even very poor density forecasts may not be detected and an incorrect forecasting model may be selected.

Proposition 2 says that the LR-tests described above (which focus on the unconditional standard deviation of an n-series) and other tests that do not cover the possibility of heteroskedasticity will tend to have no power to detect incorrect density forecasts if the forecasting model correctly specifies the conditional mean of the target variable but the forecaster incorrectly assumes normally distributed density forecasts based on the unconditional standard deviation. Propositions 1 and 2 further imply that in such situations attempts to refine the alternative hypotheses about an n-series given in (3) by including for

example various powers of a n-series or other variables will not help to detect incorrect density forecasts.

What else can be said about the properties of an n-series under a misspecified forecasting model? It can be shown (Diebold, Hahn and Tay, 1999, proposition 1) that a z-series keeps the iid property but is not uniformly distributed anymore if a sequence of true conditional densities $f(x_t \mid I_{t-1})$ belongs to a location-scale family (i.e. $X_t = \mu(I_{t-1}) + \sigma(I_{t-1})Y_t$ is an affine transformation of a random variable Y_t with a distribution D, independent of the information I_{t-1}, $\sigma(I_{t-1}) > 0$) and the forecasted densities $p(x_t \mid I_{t-1})$ adequately capture the dynamics of the first two conditional moments but belong to another location-scale family. It can easily be shown that this result extends to the corresponding n-series. It can also be demonstrated (Berkowitz, 2000, proposition 2) that if $h(n_t)$ is the density of n_t generated under the density forecast $p(x_t)$ and $\phi(n_t)$ is the standard normal density, then $\log[f(x_t)/p(x_t)] = \log[h(n_t)/\phi(n_t)]$, which implies that deviations of a density forecast from the true density will be preserved in the corresponding regions of a standard normal density.

Taken together, the discussion in this section suggests that a) misspecifications of a forecasting model will be preserved in the corresponding n-series and b) that density forecast evaluation procedures based on n-series should cover the possibility of conditional heteroskedasticity (i.e. incorporate higher conditional moments of an n-series) and tests about the distribution of an n-series.

3 Regression Framework

It is well known that for a random variable Y (with $E[Y^2] < \infty$) the orthogonal decomposition $Y = E(Y \mid H) + (Y - E(Y \mid H))$, where $E(Y \mid H)$ denotes the expectation of Y conditional on the information set H, is well defined relative to H (for details see Spanos, 1999, ch.15, or Karr, 1992, ch. 8). Thus the statistical generating mechanism for the first conditional moment for Y can be stated as $Y = E(Y \mid H) + u$. A similar orthogonal decomposition can be applied to Y^2 assuming that the required moments exist. If the transformed series $\{n_t\}_{t = 1,\ldots,m}$ is iid N(0,1), then the first two conditional moments take the form $E(N_t \mid H_t) = \mu = 0$ (independence) and $E(N_t^2 \mid H_t) = Var(N_t) = \sigma^2 = 1$ (conditional homoskedasticity and unit variance). The setting obviously also implies that $N_t = u_t$ must be distributed N(0,1) if the iid N(0,1) property holds.

In the context of density forecast evaluation many choices for variables $\in H_t$ are possible. For example H_t could contain various lags of an n-series as well as various powers and cross-products of an n-series, other variables of interest could also be included. The

important point is that a) the more general model which forms the alternative hypothesis contains the H_0 of iid N(0,1) as a special case and is based on a set of internally consistent probabilistic assumptions and that b) the more general model covers important departures from iid N(0,1) that are interesting for the purpose of density forecast evaluation. In the light of the results from the last section, the two regression functions are specified as

(4a) $$n_t = \beta_0 + \beta_1 n_{t-1} + ... + \beta_k n_{t-k} + u_t$$

(4b) $$n_t^2 = \gamma_0 + \gamma_1 n_{t-1}^2 + ... + \gamma_s n_{t-s}^2 + v_t$$

where $\{u_t\}$ and $\{v_t\}$ are martingale difference sequences (i.e. non-autocorrelated with zero expectation conditional on it's own past). In this framework the hypotheses of an iid N(0,1) n-series implies the restrictions $\beta_0 = \beta_1 = ... = \beta_k = 0$ (zero mean and independence) and $n_t \sim N(0,1)$ (normal distribution with mean zero and unit unconditional variance) in (4a) and $\gamma_0 = 1$, $\gamma_1 = ... = \gamma_s = 0$ (constant conditional unit variance, i.e. conditional homoskedasticity) in (4b). Note that for k = 1 equation (4a) is similar to the alternative hypothesis of the LR-methodology defined in equation (3) above, but there are also important differences. In contrast to (3), model (4a) accommodates conditional heteroskedasticity and does not assume normality. Hence, the model given by (4a) is more general than model (3) and includes it as a special case. In addition, equation (4b) incorporates second order dependence of n_t explicitly and includes the possibility of conditional heteroskedasticity. Hence, a test of the restriction $\gamma_1 = ... = \gamma_s = 0$ can be interpreted as an ARCH test. The restrictions on the coefficients in (4a) and (4b) can easily be tested using heteroskedasticity consistent Wald tests. Under the assumptions made in (4a) and (4b) these tests can be justified asymptotically (for details, see Hayashi, ch. 2). A joint Wald test of all β and γ restrictions under the possibility of heteroskedasticity can be carried out using standard system estimation methods.[1]

As discussed in section 2, detected deviations of an n-series from normality indicate problems with the distributional assumptions on which the density forecasts are built and tests concerning the normality of an n-series are therefore essential for the proper selection of a forecasting model. In principal one could extend the regression framework by including autoregressions of third and fourth powers of an n-series and run a joint Wald test on the restrictions implied by iid N(0,1) in the enlarged system of equations. If only a

---

[1] In large samples and in absence of heteroskedasticity the test results obtained from t- and F-tests in (4a) will be virtually identical to the results obtained from LR-tests based on (3a) since then the tests are asymptotically equivalent. In the regression framework the equivalent to the joint LR-test based on (3b) is a test of the restriction $\beta_0 = \beta_1 = 0$, $\gamma_0 = 1$, in the sub system $n_t = \beta_0 + \beta_1 n_{t-1} + u_t$, $n_t^2 = \gamma_0 + v_t$.

few restrictions are violated, a joint Wald test of this type is likely to have low power for typical sample sizes, however. One strategy is to test the relevant set of restrictions on each equation individually to explore possible directions of misspecifications more closely. This is done in the empirical applications as a further step if normality of an n-series is rejected by a Jarque-Bera normality test (Jarque and Bera, 1980) and separate tests about skewness and kurtosis in the first step of the analysis.

4 Simulation Study

This section explores the power and size of LR-tests based on (3b), joint Wald tests (W) on the system (4a) and (4b) and the JB-test for a data generating processes that is realistic for financial return series. The data generating mechanism is specified to be a GARCH(1,1) process of the form

(5) $$y_t = (h_t)^{1/2}[\nu/(\nu-2)]^{-1/2} t_\nu$$

$$h_t = \alpha_0 + \alpha_1 y_{t-1}^2 + \alpha_2 h_{t-1}$$

with zero conditional mean and innovations drawn from a fat-tailed t-distribution with ν = 5 degrees of freedom. This process displays the typical features often found in financial time series, namely conditional heteroskedasticity, fat-tailed conditional distributions and fat-tailed unconditional distributions.

The data generating process is investigated for four different parameter vectors of the variance equation. In model 1 and model 2 the parameters α_0 and α_1 are set to 0.004 and 0.06, respectively. The processes differ in their persistence parameter α_2 of the conditional variance. In model 1, α_2 is set to 0.75 which is a value closer to the lower end of the range of persistence parameters typically found for financial return series, whereas model 2 assumes α_2 = 0.90 which is a more typical value. However, empirical studies sometimes report estimates for α_2 close to one. Model 3 and model 4 take this findings into account by assuming α_1 = 0.03 and α_2 = 0.95 and α_1 = 0.01 and α_2 = 0.98, respectively.

The power of the different tests are investigated under two alternative scenarios. The first case (qml) corresponds to proposition 1 and assumes that the forecaster correctly specifies the functional form of the econometric model, but estimates the GARCH(1,1) model under the wrong assumption of gaussian innovations (i.e. performs quasi maximum likelihood estimation of the model) and hence issues normally distributed density forecasts instead of fat-tailed t-distributed forecasts. The second case (uc. normal) corresponds to proposition 2

and assumes normally distributed density forecasts based on the unconditional standard deviation of y_t, thereby wrongly neglecting conditional heteroskedasticity in addition to the incorrect distributional assumption. Each experiment is based on 10000 simulations and the rejection ratios of the different test statistics applied to n_t series resulting from one step ahead density forecasts are calculated for sample sizes of 200, 500, 1000 and 1500 observations. To investigate the size of the test statistics, the rejection ratios are also calculated for the correct models.

Table 1 reports the results of the simulation experiments. Consistent with the implications of proposition 1, the LR-test and the joint W-test (based on the first lag of n_t in (4a) and the first six lags of n_t^2 in (4b)) have virtually no power in detecting incorrect density forecasts under the qml scenario for all four models and all sample sizes. For example, in the qml scenario for model 3 (with a significance level $\alpha = 0.05$) the power of the LR-test is ranging between 0.031 for a sample size of 200 and 0.022 for the largest sample size of 1500 observations and is therefore extremely low. The same is true for the W-test under the same scenario. It's power is only between 0.031 for a sample size of 200 and 0.039 for a sample size of 1500 observations. Note that the JB tests are quite powerful (about 0.74 to 1.0) in all cases, however, indicating significant deviations from normality and therefore incorrect density forecasts. Without the additional JB-tests which virtually always reject normality, one would nearly always accept the wrong model and hence deficient density forecasts.

The simulation results suggest that the LR-test also has no power in the four models under the alternative uc. normal scenario, as predicted by proposition 2. For example, looking at model 3 again (significance level $\alpha = 0.05$) the power of the LR-test is again very low and only in a range of 0.026 to 0.038. In contrast to the results for the LR-test, the joint W-test (which includes a test for conditional heteroskedasticity) has now reasonable power in detecting incorrect density forecasts for sample sizes of 1000 (power = 0.411) and 1500 observations (power = 0.543). With respect to the different persistence parameters, the simulations under the scenario uc.normal show that the W-test tends to have reasonable power for $\alpha_2 = 0.75$, $\alpha_2 = 0.90$ and $\alpha_2 = 0.95$ and sample sizes of 1000 or more observations. However, the power of the W-test decreases rather sharply for $\alpha_2 = 0.98$, suggesting that additional distributional tests are important for successful density forecast evaluations. The simulation results for uc. normal again highlight this point. Like in the qml scenario, the JB-test rejects normality for all four models with rejection rates between 0.84 to 1.0, thereby correctly indicating deficient density forecasts most of the time.

Table 1. Power and size of density forecast evaluation methodologies based on n-series from simulated GARCH-t models

model 1 $y_t = \sqrt{h_t} * t_5$, $h_t = 0.004 + 0.06y_{t-1} + 0.75h_{t-1}$

	LR		W		JB	
	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.05$
power: qml						
t = 200	0.057	0.028	0.053	0.033	0.798	0.758
t = 500	0.050	0.024	0.058	0.038	0.992	0.989
t = 1000	0.040	0.020	0.053	0.033	1.000	1.000
t = 1500	0.044	0.020	0.055	0.035	1.000	1.000
power: uc. normal						
t = 200	0.067	0.035	0.204	0.161	0.894	0.864
t = 500	0.070	0.040	0.348	0.289	0.997	0.995
t = 1000	0.071	0.039	0.516	0.446	1.000	1.000
t = 1500	0.077	0.042	0.652	0.583	1.000	1.000

model 2 $y_t = \sqrt{h_t} * t_5$, $h_t = 0.004 + 0.06y_{t-1} + 0.90h_{t-1}$

power: qml						
t = 200	0.052	0.023	0.057	0.033	0.796	0.756
t = 500	0.049	0.023	0.059	0.035	0.992	0.989
t = 1000	0.045	0.022	0.060	0.038	1.000	1.000
t = 1500	0.040	0.017	0.059	0.039	1.000	1.000
power: uc. normal						
t = 200	0.064	0.033	0.273	0.216	0.893	0.865
t = 500	0.077	0.041	0.540	0.473	0.998	0.996
t = 1000	0.088	0.050	0.800	0.749	1.000	1.000
t = 1500	0.095	0.056	0.922	0.890	1.000	1.000

model 3 $y_t = \sqrt{h_t} * t_5$, $h_t = 0.004 + 0.03y_{t-1} + 0.95h_{t-1}$

power: qml						
t = 200	0.061	0.031	0.052	0.031	0.782	0.740
t = 500	0.051	0.028	0.060	0.038	0.993	0.989
t = 1000	0.052	0.023	0.060	0.038	1.000	1.000
t = 1500	0.044	0.022	0.059	0.039	1.000	1.000
power: uc. normal						
t = 200	0.053	0.026	0.146	0.105	0.875	0.842
t = 500	0.060	0.029	0.288	0.227	0.997	0.996
t = 1000	0.062	0.032	0.476	0.411	1.000	1.000
t = 1500	0.067	0.038	0.609	0.543	1.000	1.000

model 4 $y_t = \sqrt{h_t} * t_5$, $h_t = 0.004 + 0.01y_{t-1} + 0.98h_{t-1}$

power: qml						
t = 200	0.067	0.033	0.050	0.029	0.786	0.744
t = 500	0.062	0.032	0.047	0.030	0.993	0.988
t = 1000	0.059	0.034	0.049	0.033	1.000	1.000
t = 1500	0.061	0.034	0.053	0.034	1.000	1.000
power: uc. normal						
t = 200	0.050	0.023	0.081	0.055	0.882	0.846
t = 500	0.046	0.022	0.107	0.075	0.997	0.994
t = 1000	0.041	0.020	0.146	0.112	1.000	1.000
t = 1500	0.053	0.027	0.185	0.142	1.000	1.000

Table 1. (continued): Power and size of density forecast evaluation methodologies based on n-series from simulated GARCH-t models

---

model 1 $y_t = \sqrt{h_t} * t_5$, $h_t = 0.004 + 0.06y_{t-1} + 0.75h_{t-1}$

	LR		W		JB	
	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.05$
size:						
t = 200	0.100	0.052	0.093	0.049	0.078	0.045
t = 500	0.098	0.049	0.099	0.051	0.090	0.049
t = 1000	0.099	0.051	0.103	0.051	0.092	0.048
t = 1500	0.100	0.052	0.100	0.051	0.091	0.047

model 2 $y_t = \sqrt{h_t} * t_5$, $h_t = 0.004 + 0.06y_{t-1} + 0.90h_{t-1}$

size:						
t = 200	0.102	0.052	0.099	0.055	0.079	0.047
t = 500	0.100	0.049	0.100	0.051	0.090	0.048
t = 1000	0.101	0.052	0.099	0.051	0.092	0.046
t = 1500	0.096	0.047	0.095	0.048	0.097	0.051

model 3 $y_t = \sqrt{h_t} * t_5$, $h_t = 0.004 + 0.03y_{t-1} + 0.95h_{t-1}$

size:						
t = 200	0.102	0.050	0.092	0.050	0.080	0.044
t = 500	0.102	0.050	0.095	0.049	0.084	0.046
t = 1000	0.101	0.051	0.097	0.049	0.095	0.051
t = 1500	0.099	0.051	0.102	0.055	0.092	0.050

model 4 $y_t = \sqrt{h_t} * t_5$, $h_t = 0.004 + 0.01y_{t-1} + 0.98h_{t-1}$

size:						
t = 200	0.098	0.049	0.096	0.054	0.075	0.043
t = 500	0.104	0.053	0.101	0.055	0.089	0.049
t = 1000	0.101	0.050	0.100	0.051	0.090	0.047
t = 1500	0.100	0.052	0.100	0.050	0.096	0.046

Notes: For all simulated GARCH-t models t_5 denotes a t-distributed random variable with mean zero and five degrees of freedom, h_t denotes the conditional variance and y_t stands for the generated returns. LR is the short cut for a joint likelihood ratio test as defined in (3b) of zero mean, zero correlation and unit variance of an n-series derived from the model. W denotes a joint Wald test of an n-series for iid N(0,1) as implied by the system (4a) and (4b). JB denotes the Jarque-Bera test statistic. The acronym qml indicates that the n-series on which the different tests are carried out are derived from quasi maximum likelihood estimates (i.e. conditionally normally distributed density forecasts) of the model, uc. normal indicates that the n-series from the model is generated under the assumption of unconditionally normally distributed density forecasts. For all models $y_1 = 0$ and the implied unconditional variance are used as starting values in the simulations.

With respect to the size of the different test statistics the simulations show that the LR and the W-test have virtually always the correct size for all models and sample sizes. The size of the JB-test is found to be slightly too low for the sample sizes considered. Taken together, the results of the simulation experiments suggest that the regression based density forecast evaluation methodology in conjunction with normality tests is a quite powerful tool for the analysis of density forecasts and model specifications.

5 Data and Forecasting Models

The analysis of the density forecasts from the forecasting models outlined below is based on daily time series of the S&P 500, DAX and ATX stock market indices. The data set obtained from Datastream covers the period from 1/26/1990 to 1/26/2000 and contains 2,609 observations per index. Daily logarithmic returns are calculated as $x_t = \ln(P_t) - \ln(P_{t-1})$ where P_t denotes the level of the index at day t.

One-step-ahead density forecasts of daily returns are generated from seven popular models. The first model is an equally weighted moving average (MA) of squared returns with a rolling time window of 250 trading days. The MA forecast of the variance of a return at time t is given by

$$(5) \qquad \sigma_t^2 = \sum_{i=t-n}^{t-1} x_i^2 / n .$$

The second model is the exponentially weighted moving average (EWMA) of squared returns with a smoothing parameter $\lambda = 0.94$ as proposed by J.P. Morgan.[2]

$$(6) \qquad \sigma_t^2 = (1-\lambda) x_{t-1}^2 + \lambda \sigma_{t-1}^2 .$$

In (5) and (6) it is assumed that the mean of the daily returns are approximately zero.[3] It is further assumed that the returns are conditionally normal with variance σ_t^2. Therefore, both models imply normal density forecasts with mean zero based on the variances generated from (5) and (6), respectively.

---

2 For further details, see RiskMetrics™ (1996).

3 This assumption is often made in practical applications of MA and EWMA models because it is argued that incorporation of the rather imprecise estimates of the mean of a daily return series (which are often close to zero) tend to produce inferior volatility predictions. For a discussion of this issue, see Figlewski (1994).

The next four forecasting models are all variants of GARCH(1,1) models. In contrast to MA and EWMA specifications, which can be applied to squared returns directly, the coefficients of GARCH models must be estimated with maximum likelihood methods. For all GARCH models the equation for the conditional mean is specified as an AR(1) process

(7) $$x_t = \omega_0 + \omega_1 x_{t-1} + \eta_t$$

to capture aggregation effects and other sources that might induce correlation into a return series. The dynamics of the conditional variances are specified as

(8a) $$h_t = \alpha_0 + \alpha_1 \eta_{t-1}^2 + \alpha_2 h_{t-1}$$

(8b) $$h_t = \alpha_0 + \alpha_1 \eta_{t-1}^2 + \gamma \eta_{t-1}^2 d_{t-1} + \alpha_2 h_{t-1}$$

$$d_t = \begin{cases} 1 & if\ \eta_t \prec 0 \\ 0 & otherwise \end{cases}.$$

Variant (8a) is the standard GARCH (1,1) specification (Bollerslev, 1986) where positive and negative innovations are treated symmetrically. Specification (8b) is the GARCH model proposed in Glosten, Jagannathan and Runkle (1993), which allows for asymmetric reactions to news on the stock market.

Equation (7) together with (8a) or (8b) determine the location and shape of the density forecasts from GARCH models. In equation (7) the coefficients ω_0 and ω_1 determine the conditional mean of the return x_t and hence the location of a density forecast at time t and the coefficients in (8a) or (8b) specify the dynamics and the size of the conditional second moments of the forecasts. The distributional form of the density forecasts is given by the distribution assumed for the disturbance term η_t. In the empirical applications the GARCH-models are estimated under the assumption of normally distributed errors and under the assumption of t-distributed error terms. In each application of the t-distribution, the degrees of freedom parameter of the t-distribution is estimated jointly with the other model coefficients. The reason for assuming a Student-t distribution is that although in the GARCH framework conditionally normal distributions produce fat-tailed unconditional distributions, often not all of the excess kurtosis is captured under the assumption of conditional normality. Since the Student's-t distribution is able to produce (symmetric) fat-tailed conditional densities, forecasts based on the Student-t distribution might be better able to capture excess kurtosis in the data.

The last model is the scaled Student's t distribution, for which the density is given by

$$(9) \qquad f(\eta_t) = \frac{\Gamma\left[\frac{1}{2}(\nu+1)\right]}{\Gamma\left[\frac{1}{2}\nu\right]}(\sigma^2\nu\pi)^{-\frac{1}{2}}\left(1+\frac{\eta_t^{\,2}}{\sigma^2\nu}\right)^{-\frac{(\nu+1)}{2}}$$

with expectation $E(\eta_t) = 0$ and variance $Var(\eta_t) = \sigma^2(\nu/\nu\text{-}2)$. In (8) $\Gamma(.)$ represents the gamma function, ν is the degrees of freedom parameter and σ^2 denotes the scale parameter. We allow for a time dependent first moment of the density forecasts since (9) is applied to the residuals obtained from the mean equation (6). Hence the location of the density forecasts based on (9) can change over time, but the shape of the forecasted densities remains the same (i.e. constant conditional variance is assumed). The intention behind this model is to analyze the consequences of neglected second moment dynamics if a an unconditional fat-tailed distribution is already assumed. The properties of the resulting density forecasts should provide valuable information about the relative importance of distributional assumptions versus assumptions about the dynamics of second moments.

6 Empirical Results

It is interesting how well the individual models perform in-sample as well as out-of-sample. Therefore, the data available for each daily index return series are divided into two subsamples. The first sample (1/29/1990 to 1/26/1996), contains 1,564 observations and is reserved for the estimation of the various GARCH models, the scaled t distributions and for the in-sample evaluation of the density forecasts. The remaining 1,044 observations of the data set, covering the period from 1/29/1996 to 1/26/2000, are used to evaluate out-of-sample density forecasts. The density forecasts of the MA models are based on a rolling window of 250 trading days shifted each day. EWMA density forecasts are obtained from the recursive expression (6). The in-sample density forecasts from the GARCH models are based on parameters estimated from the in-sample period data. The out-of-sample density forecasts are based on coefficients updated once a year using a sample of fixed length containing the latest 1564 observations available at the time of updating. The parameters for the scaled t distributions are estimated from the in-sample period data and both the in- and out-of-sample density forecasts are based on these parameters.

Table 2 provides a summary statistic on each daily return series for both samples.

Table 2. Summary statistics of Return Series

	in-sample period 1/14/1991-1/26/1996			out-of-sample period 1/29/1996-1/26/2000		
	S&P 500	**DAX**	**ATX**	**S&P 500**	**DAX**	**ATX**
mean	0.000523	0.000461	9.56E-05	0.000778	0.001005	6.67E-05
maximum	0.036642	0.072875	0.076139	0.049887	0.061057	0.052623
minimum	-0.037272	-0.098707	-0.074695	-0.071127	-0.083822	-0.086995
std. dev	0.006373	0.009683	0.011174	0.010891	0.014343	0.011651
skewness	0.056141	-0.475196	0.373809	-0.482427	-0.581209	-0.907929
kurtosis	5.908284	14.41501	10.75624	7.638396	6.321387	8.761885
Jarque-Bera	463.7719	7183.514	3324.317	974.5155	537.6211	1584.563

The summary statistics indicate that all return series display a significant amount of excess kurtosis (the kurtosis of a normal distribution is 3) in both samples. Hence, all unconditional distributions have fatter tails than the normal distribution, which implies that extreme events tend to occur more frequently than a normal distribution would predict. This result is typical for most financial time series. Note that all return distributions over the out-of-sample period show greater negative skewness than over the in-sample period.

Tables 3, 4 and 5 report the in-sample and out-of-sample evaluation results about the quality of the one-step-ahead density forecasts generated by the different models. In the tables LR denotes the joint likelihood ratio test LR_2 of correct density forecasts given by (3b) and W_j denotes a joint test of the restriction $\beta_0 = \beta_1 = \gamma_1 = ... = \gamma_6 = 0$, $\gamma_0 = 1$ in the system given by (4a) and (4b) with the first lag of n_t and the first six lags of n_t^2 under the possibility of heteroskedasticity. Estimated coefficients and p-values from individual t-tests for zero β coefficients are reported under β_0 and β_1. These estimates and tests come from regressions (4a) under the assumption of homoskedasticity. Because of the large sample size and the assumption of homoskedasticity the reported p-values of the t-statistics are virtually identical to the p-values from corresponding individual LR-tests and hence directly comparable. Under σ chi-square tests of the hypothesis of an unconditional unit variance of an n-series are reported and J-B and ARCH-F denote Jarque-Bera normality tests of an n-series and F-tests for conditional homoskedasticity in regressions (4b).

Table 3. Test statistics of n-series of density forecasts of daily S&P500 stock market returns

	MA-N	EWMA-N	GARCH-N	GJR-N	GARCH-t	GJR-t	Scaled-t
	in-sample period 1/14/1991 – 1/26/1996						
LR	13.314 (0.004)	17.391 (0.001)	0.675 (0.879)	0.828 (0.843)	3.926 (0.270)	2.656 (0.448)	16.399 (0.001)
$\mathbf{W_j}$	29.570 (0.001)	13.647 (0.135)	1.619 (0.996)	2.316 (0.985)	9.421 (0.421)	10.143 (0.339)	59.272 (0.000)
β_0	0.076 (0.004)	0.081 (0.006)	0.002 (0.951)	0.013 (0.620)	0.013 (0.615)	0.008 (0.780)	0.013 (0.606)
β_1	0.006 (0.817)	0.035 (0.205)	-0.012 (0.671)	-0.017 (0.537)	-0.032 (0.248)	-0.027 (0.327)	-0.056 (0.044)
σ^2	0.911 (0.009)	1.104 (0.005)	0.974 (0.252)	0.983 (0.334)	0.942 (0.066)	0.952 (0.107)	0.871 (0.000)
J-B	261.91 (0.000)	438.18 (0.000)	353.91 (0.000)	386.25 (0.000)	0.075 (0.963)	0.21 (0.898)	0.60 (0.741)
ARCH-F	2.946 (0.007)	0.231 (0.967)	0.161 (0.987)	0.242 (0.962)	0.754 (0.606)	1.167 (0.321)	6.209 (0.000)
	out-of-sample period 1/29/1996 – 1/26/2000						
LR	29.815 (0.000)	15.718 (0.001)	21.893 (0.000)	38.574 (0.000)	21.532 (0.000)	35.531 (0.000)	213.885 (0.000)
$\mathbf{W_j}$	49.769 (0.000)	13.887 (0.126)	13.272 (0.151)	12.971 (0.164)	26.258 (0.002)	42.946 (0.000)	201.513 (0.000)
β_0	0.079 (0.021)	0.063 (0.056)	0.010 (0.764)	0.021 (0.557)	0.031 (0.368)	0.028 (0.415)	0.061 (0.143)
β_1	0.028 (0.367)	0.052 (0.091)	0.008 (0.785)	0.002 (0.955)	0.044 (0.155)	0.051 (0.103)	-0.004 (0.906)
σ^2	1.221 (0.000)	1.132 (0.002)	1.219 (0.000)	1.297 (0.000)	1.201 (0.000)	1.270 (0.000)	1.778 (0.000)
J-B	878.01 (0.000)	553.44 (0.000)	699.38 (0.000)	547.63 (0.000)	10.51 (0.005)	12.54 (0.002)	20.25 (0.000)
ARCH-F	6.253 (0.000)	0.669 (0.675)	1.208 (0.299)	0.375 (0.895)	1.007 (0.419)	1.799 (0.096)	7.315 (0.000)

Notes: P-values in parenthesis

Table 4. Test statistics of n-series of density forecasts of daily DAX stock market returns

	MA-N	EWMA-N	GARCH-N	GJR-N	GARCH-t	GJR-t	Scaled-t
	in-sample period 1/14/1991 – 1/26/1996						
LR	7.219	37.772	2.675	2.122	3.100	3.137	16.088
	(0.065)	(0.000)	(0.444)	(0.547)	(0.376)	(0.371)	(0.001)
W_j	30.428	6.245	0.945	1.010	10.421	11.732	80.725
	(0.000)	(0.715)	(0.999)	(0.999)	(0.317)	(0.229)	(0.000)
β_0	0.041	0.035	0.005	0.019	0.023	0.023	0.029
	(0.125)	(0.250)	(0.859)	(0.486)	(0.388)	(0.397)	(0.261)
β_1	0.042	0.049	-0.012	-0.005	0.022	0.021	0.006
	(0.131)	(0.072)	(0.671)	(0.867)	(0.417)	(0.436)	(0.818)
σ^2	0.934	1.241	0.941	0.952	0.951	0.949	0.857
	(0.060)	(0.000)	(0.062)	(0.106)	(0,102)	(0.094)	(0.000)
J-B	1171.01	56697.15	20615.00	19143.43	2.05	1.81	0.04
	(0.000)	(0.000)	(0.000)	(0.000)	(0.358)	(0.404)	(0.981)
ARCH-F	4.053	0.036	0.054	0.039	1.164	1.380	9.520
	(0.000)	(0.999)	(0.999)	(0.999)	(0.323)	(0.219)	(0.000)
	out-of-sample period 1/29/1996 – 1/26/2000						
LR	27.463	15.238	19.880	28.404	14.200	15.780	118.587
	(0.000)	(0.002)	(0.000)	(0.000)	(0.003)	(0.001)	(0.000)
W_j	106.456	14.727	21.682	21.667	17.508	19.118	280.119
	(0.000)	(0.099)	(0.010)	(0.010)	(0.041)	(0.024)	(0.000)
β_0	0.094	0.100	0.045	0.054	0.069	0.069	0.109
	(0.006)	(0.007)	(0.187)	(0.121)	(0.037)	(0.039)	(0.005)
β_1	-0.026	-0.008	-0.000	-0.002	0.013	0.015	-0.001
	(0.397)	(0.809)	(0.989)	(0.938)	(0.679)	(0.635)	(0.965)
σ^2	1.199	1.121	1.197	1.239	1.138	1.150	1.524
	(0.000)	(0.004)	(0.000)	(0.000)	(0.001)	(0.001)	(0.000)
J-B	497.75	232.47	140.22	154.81	19.10	18.14	19.76
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
ARCH-F	15.277	0.641	1.909	1.361	0.645	0.709	29.059
	(0.000)	(0.698)	(0.076)	(0.227)	(0.694)	(0.642)	(0.000)

Notes: P-values in parenthesis

Table 5. Test statistics of n-series of density forecasts of daily ATX stock market returns

			in-sample period 1/14/1991 – 1/26/1996				
	MA-N	**EWMA-N**	**GARCH-N**	**GJR-N**	**GARCH-t**	**GJR-t**	**Scaled-t**
LR	46.595 (0.000)	71.368 (0.000)	2.793 (0.425)	1.841 (0.606)	3.137 (0.373)	3.293 (0.349)	21.939 (0.000)
$\mathbf{W_j}$	111.679 (0.000)	68.190 (0.000)	7.010 (0.636)	9.267 (0.413)	8.088 (0.325)	8.742 (0.461)	190.188 (0.000)
β_0	0.001 (0.956)	-0.019 (0.506)	-0.006 (0.833)	0.017 (0.519)	0.001 (0.980)	-0.000 (0.996)	0.012 (0.649)
β_1	0.176 (0.000)	0.202 (0.000)	-0.010 (0.708)	-0.008 (0.779)	-0.019 (0.499)	-0.019 (0.496)	-0.053 (0.056)
σ^2	0.914 (0.012)	1.162 (0.000)	0.939 (0.057)	0.956 (0.128)	0.938 (0.055)	0.937 (0.050)	0.844 (0.000)
J-B	902.64 (0.000)	4692.83 (0.000)	4356.85 (0.000)	2788.45 (0.000)	0.079 (0.961)	0.040 (0.980)	0.880 (0.644)
ARCH-F	11.214 (0.000)	1.548 (0.159)	1.015 (0.414)	1.191 (0.308)	0.927 (0.475)	0.811 (0.561)	26.090 (0.000)
			out-of-sample period 1/29/1996 – 1/26/2000				
LR	6.210 (0.102)	16.127 (0.001)	4.262 (0.235)	4.464 (0.215)	2.039 (0.564)	1.646 (0.649)	34.580 (0.000)
$\mathbf{W_j}$	117.892 (0.000)	21.880 (0.009)	9.805 (0.367)	7.868 (0.547)	6.216 (0.718)	5.443 (0.794)	295.028 (0.000)
β_0	0.017 (0.606)	0.019 (0.556)	0.000 (0.994)	0.015 (0.635)	0.021 (0.502)	0.018 (0.565)	0.042 (0.163)
β_1	0.038 (0.220)	0.094 (0.002)	-0.064 (0.039)	-0.064 (0.040)	-0.029 (0.342)	-0.023 (0.453)	-0.175 (0.000)
σ^2	1.096 (0.017)	1.116 (0.005)	0.999 (0.493)	1.007 (0.429)	1.038 (0.193)	1.039 (0.183)	0.965 (0.212)
J-B	2419.55 (0.000)	353.21 (0.000)	161.35 (0.000)	125.51 (0.000)	9.24 (0.010)	8.63 (0.013)	6.90 (0.032)
ARCH-F	19.190 (0.000)	1.600 (0.144)	0.924 (0.477)	0.568 (0.756)	0.692 (0.656)	0.629 (0.707)	43.289 (0.000)

Notes: P-values in parenthesis

The tests of the n-series for the simple MA and the EWMA models indicate a rather poor performance in-sample as well as out-of-sample. The W_j tests, which in contrast to the LR_2-test, also cover the dynamics of the conditional second moments do not always reject the hypotheses of correct density forecasts for the EWMA-models because due to the similarity with GARCH-models, EWMA-models often provide a good approximation of the volatility dynamics which leads to more frequent non-rejections. In fact, all n-series from the EWMA-models pass the separate ARCH-F tests. However, all n-series generated from MA and EWMA models clearly do not pass the J-B normality test, as indicated by the rather large values of the Jarque-Bera test statistics. In addition, the individual t-tests sometimes indicate problems with the location and the dynamics of the density forecasts.

The results for the GARCH- and GJR-models with normally distributed errors clearly highlight the danger of using only LR tests without additional tests for normality. The LR tests support the hypotheses of correct density forecasts all times over the in-sample period. The J-B normality tests, however, strongly reject normality in all cases indicating severe problems with the assumption of normally distributed density forecasts. Without the additional tests for normality one would have incorrectly accepted all GARCH-models with normally distributed errors over the in-sample period. Without normality tests the W_j statistic alone would of course also lead to incorrect conclusions. In conjunction with normality tests, however, the results do not support normally distributed density forecasts from GARCH models. Over the out-of-sample period the results are somewhat mixed. The LR tests reject, except for the ATX where the W_j statistics also accept due to the absence of ARCH-effects. The Wj statistics also weakly support the GARCH-n and GJR-n models for the S&P 500. However, the additional J-B tests strongly reject normality again.

The results for GARCH- and GJR-models with t-distributed errors are quite different from the models with normally distributed errors. Both models pass all tests over the in-sample-period indicating good density forecasts. In the case of the S&P 500 and the DAX, the GARCH-t and GJR-t models are not supported by the LR and W_j statistics over the out-of sample period and the J-B normality test rejects normality at conventional significance levels in all three cases. However, the value of the J-B test statistic is small and by far lower, compared to the models that assume normally distributed density forecasts.[4]

[4] Another interesting point is that the incorporation of an asymmetric reaction of volatility to positive and negative innovations into the econometric specification does not seem to be crucial for the purpose of density forecasting, although we found some evidence for significant positive γ coefficients for the GJR models implying a larger impact of negative innovations on volatility.

Individual chi-square statistics for skewness SK = 0 and kurtosis K = 3 for the out-of-sample n-series of the GARCH-t and GJR-t models reported in table 5 provide additional insights about likely directions of misspecification. The statistically significant negative skewness coefficients for the n-series suggests that the main deficiency of the density forecasts from these models might result from the symmetry imposed by the t-distribution. Additional F-tests (F) of the restriction $\delta_1 = \delta_2 = ... = \delta_5 = 0$ from the regression $n^3_t = \delta_0 + \delta_1 n^3_{t-1} + . . . + \delta_5 n^3_{t-5} + e_t$ of the cubed n's on it's first five lags provide information about time dependence of skewness of the n-series. If there is no time dependence, than the lagged cubed n_t's should not help to predict actual cubed n_t's. The F-tests does not reject the hypotheses of time independent skewness for the transformed DAX and ATX series. The F-tests for n^3-series from the GARCH-t and GJS-t models for the S&P 500 series indicate time dependent skewness. Density forecasts from models along the lines of Hansen (1994) that allow for time dependent skewness might therefore be more appropriate for the S&P 500. Such models are beyond the scope of this paper, however.

Table 6. Skewness and kurtosis of transformed stock market returns, 1/29/1996 – 1/26/2000

		K	SK	F-test
S&P500	**GARCH-t**	2.699 (0.041)	-0.191 (0.012)	4.031 (0.001)
	GJR-t	2.633 (0.016)	-0.196 (0.010)	3.793 (0.002)
DAX	**GARCH-t**	2.843 (0.300)	-0.322 (0.000)	0.857 (0.509)
	GJR-t	2.850 (0.323)	-0.315 (0.000)	0.794 (0.554)
ATX	**GARCH-t**	2.839 (0.289)	-0.216 (0.004)	1.587 (0.161)
	GJR-t	2.808 (0.206)	-0.201 (0.008)	1.426 (0.212)

Notes: P-values in parenthesis

The last model to be discussed is the scaled t-distribution with a constant conditional second moment. This model, neglecting the time dependence in the conditional second moments, is always strongly rejected by the LR and Wj statistics, although the J-B-statistics looks good in all cases, often supporting normality. The ARCH-F tests clearly indicate serious heteroskedasticty. A comparison of the test results for the GARCH-t and GJR-t models with scaled t-distributions shows that both, proper distributional assumptions

and a reasonable model of the dynamics of the relevant conditional moments are necessary to obtain good density forecasts.

7 Conclusions

Based on the fact that correct density forecasts for a stochastic process imply iid N(0,1) data under certain transformations of the realizations of a process with respect to the corresponding predicted conditional densities, a simple regression framework in conjunction with normality tests was proposed to evaluate the quality of density forecasts obtained from econometric time series models. The methodology is not only useful to examine the quality of density forecasts per se, because it is also applicable to identify the nature of misspecifications of the forecasting model being used. It was further demonstrated theoretically, as well as in simulation experiments and in empirical applications that likelihood ratio tests focusing only on the mean, correlation and unconditional variance of a transformed series may lead to misleading conclusions about the quality of density forecasts and the associated forecasting models if no additional normality- and heteroskedasticity tests are conducted.

The empirical results about the quality of in- and out-of-sample one-step-ahead density forecasts of daily returns from the S&P 500, DAX and ATX suggest that GARCH-models with t-distributed errors are able to produce good density forecasts over the in-sample period. Experiments with unconditional t-distributions (thereby ignoring the dynamics in the second moments) show that the choice of a fat-tailed distribution alone is not enough to obtain acceptable density forecasts. Distributional assumptions as well as the correct specification of conditional moments play an important role. The performance of GARCH-t and GJR-t models is weaker out-of-sample, but still better compared to the other models. Separate skewness- and kurtosis tests and an analysis of the correlation structure in the third conditional moments of the transformed series indicates that GARCH-models with skewed fat-tailed conditional distributions might be more appropriate to describe the return series over the out-of sample period. In the case of the S&P 500, skewness was also found to be time varying. Extensions of statistical models of financial returns to higher order conditional moments beyond the conditional variance might therefore be an interesting direction for future research.

Appendix

Proof of Proposition 1:

The random variable X_t in it's standardized form is given by $S_t = (X_t - \mu(I_{t-1}))/\sigma(I_{t-1})$ and the probability integral transformation (1) can be written as $Z_t = P_t(S_t)$ where $P_t(.)$ is the assumed distribution function of the density forecasts. The n-transformation applied to S_t can then be expressed as $N_t = F_N^{-1}[P_t(S_t)]$. Since the density forecast $p(X_t | I_{t-1})$ is assumed to be a normal density it follows that $N_t = F_N^{-1}[P_t(S_t)] = F_N^{-1}[F_N(S_t)] = S_t$. The fact that the predicted densities $p(X_t | I_{t-1})$ adequately capture the first two conditional moments of the density forecasts implies that $S_t = Y_t$ for all t. But then $N_t = Y_t$ is iid D(0,1) and the result follows.

Proof of Proposition 2:

a) Assume a normally distributed density forecast for X_t based on the unconditional standard deviation σ. The random variable X_t in it's standardized form can be expressed as

(A1) $$S_t = \frac{X_t - \mu(I_{t-1})}{\sigma} = \frac{\sigma(I_{t-1})Y_t}{\sigma}$$

Since $N_t = F_N^{-1}[P_t(S_t)] = F_N^{-1}[F_N(S_t)] = S_t$ holds for a normal distribution function we can substitute N_t for S_t in (A1).

The conditional second moment of N_t is given by

(A2) $$E(N_t^2 | I_{t-1}) = E\left(\frac{\sigma(I_{t-1})^2}{\sigma^2} Y_t^2 | I_{t-1}\right)$$

$$= \frac{\sigma(I_{t-1})^2}{\sigma^2} E(Y_t^2 | I_{t-1})$$

$$= \frac{\sigma(I_{t-1})^2}{\sigma^2} E(Y_t^2).$$

Since $\sigma(I_{t-1})^2$ varies across t it follows that the second conditional moment of N_t varies across t which proofs conditionally heteroskedasticity.

b) conditional mean 0:

(B1) $$E(N_t \mid I_{t-1}) = E\left(\frac{\sigma(I_{t-1})}{\sigma} Y_t \mid I_{t-1}\right) = \frac{\sigma(I_{t-1})}{\sigma} E(Y_t \mid I_{t-1}) = \frac{\sigma(I_{t-1})}{\sigma} E(Y_t) = 0$$

uncorrelatedness:

(B2) $$E(N_t N_{t-j}) = E\left[E\left(N_t N_{t-j} \mid N_{t-j}\right)\right]$$

$$= E\left[E\left(N_t \mid N_{t-j}\right) N_{t-j}\right].$$

Since N_t has mean zero conditional on I_{t-j} and I_{t-j} of course contains N_{t-j} , it follows that $E(N_t \mid N_{t-j}) = 0$ and hence $E(N_t N_{t-j}) = 0$.

unconditional mean 0:

(B3) $E(N_t) = 0$ follows immediately from $E(Y_t) = 0$.

c) unconditional standard deviation of 1:

(C1) $$E(N_t^2) = E\left(\frac{\sigma(I_{t-1})^2}{\sigma^2} Y_t^2\right) = \frac{1}{\sigma^2} E[\sigma(I_{t-1})^2] E(Y_t^2)$$

Write the model for X_t in the form $X_t = \mu(I_{t-1}) + \varepsilon_t$, where $\varepsilon_t \sim D(0, \sigma(I_{t-1}))$. Then

(C2) $$E[\sigma(I_{t-1})^2] = E[E(\varepsilon_t^2 \mid I_{t-1})] = E(\varepsilon_t^2) = \sigma^2 .$$

Since $E(Y_t^2) = 1$ by assumption, it follows that $E(N_t^2) = 1$ which proofs point c.

References

Berkowitz, J. (2000). Testing Density Forecasts, with Applications to Risk Management. Working Paper, University of California, Irvine (forthcoming in the Journal of Business and Economic Statistics).

Bollerslev, T. (1986). Generalized Autoregressive Conditional Heteroscedasticity. Journal of Econometrics 31, 307-327.

Bollerslev, T. & Wooldridge, J. M. (1992). Quasi-Maximum Likelihood Estimation and Inference in Dynamic Models with Time-Varying Covariances. Econometric Reviews, 11(2), 143-172.

Clements, M. P., & Smith, J (2000). Evaluating the Forecast Densities of Linear and Non-linear Models: Applications to Output Growth and Unimployment. Journal of Forecasting, 19, 255-276.

Crnkovic, C., & Drachman, J. (1997). Quality Control. In VaR: Understanding and Applying Value-at-Risk. London: Risk Publications.

Diebold, F. X., Gunther, T. A., & Tay, A. S. (1998). Evaluating Density Forecasts, with Applications to Financial Risk Management. International Economic Review, **39**, 863-883.

Diebold, F. X, Tay, A. S., & Wallis, K. F. (1999). Evaluating Density Forecasts of Inflation: The Survey of Professional Forecasters. In R. Engle and H. White (eds.), Festschrift in Honor of C.W.J. Granger, 76-90. Oxford: Oxford University Press.

Diebold, F. X., Hahn, J., & Tay, A. S. (1999). Multivariate Density Forecast Evaluation and Calibration in Financial Risk Management: High-Frequency Returns on Foreign Exchange. The Review of Economics and Statistics, 81(4), 661-673.

Figlewski, S. (1994). Forecasting volatility using historical data. New York University

Working Paper, S-94-13.

Frühwirth-Schnatter, S. (1996). Recursive residuals and model diagnostics for normal and non-normal state space models. Environmental and Ecological Statistics, 3, 291-309.

Glosten, L., Jagannathan, R. & Runkle, D. (1992). On the Relation Between the Expected Value and the Volatility of Nominal Excess Returns on Stocks. Journal of Finance, 46, 1779-1801.

Hansen, B. (1994). Autoregressive Conditional Density Estimation. International Economic Review, 35, 705-792.

Hayashi, F. (2000). Econometrics. Princeton University Press.

Jarque, C. M. & Bera, A. K. (1980). Efficient Tests for Normality, Heteroskedasticity and Serial Independence of Regression Residuals. Economics Letters, 6, 255-259.

Jorion, P. (1997). Value at Risk. The New Benchmark for Controlling market Risk, Irwin.

Karr, A. F. (1993). Probability, New York, Berlin, Heidelberg: Springer- Verlag,

Kaufmann, S. (2000). Measuring business cycles with a dynamic Markov switching factor model: an assessment using Bayesian simulation methods. Econometrics Journal, 3, 39-65.

Lumsdaine, R. (1996). Consistency and Asymptotic Normality of the Quasi-Maximum Likelihood Estimator in IGARCH(1,1) and Covariance Stationary GARCH(1,1) Models. Econometrica, 64(3), 575-596.

RiskMetricsTM (1996). Technical Document, JP Morgan Global Research, New York.

Rosenblatt, M. (1952). Remarks on a Multivariate Transformation. Annals of Mathematical Statistics, 23, 470-472.

Spanos Aris (1999). Probability Theory and Statistical Inference: econometric modeling with observational data. Cambridge University Press.

Tay, A. S. & Wallis, K. F. (2000). Density forecasting: A Survey. Journal of Forecasting, 19 (4), 235-254.

The following papers have been published since 2000:

February	2000	How Safe Was the „Safe Haven“? Financial Market Liquidity during the 1998 Turbulences	Christian Upper
May	2000	The determinants of the euro-dollar exchange rate – Synthetic fundamentals and a non-existing currency	Jörg Clostermann Bernd Schnatz
July	2000	Concepts to Calculate Equilibrium Exchange Rates: An Overview	Ronald MacDonald
August	2000	Core inflation rates: A comparison of methods based on west German data	Bettina Landau
September	2000	Exploring the Role of Uncertainty for Corporate Investment Decisions in Germany	Ulf von Kalckreuth
November	2000	Central Bank Accountability and Transparency: Theory and Some Evidence	Sylvester C.W. Eijffinger Marco M. Hoeberichts
November	2000	Welfare Effects of Public Information	Stephen Morris Hyung Song Shin
November	2000	Monetary Policy Transparency, Public Commentary, and Market Perceptions about Monetary Policy in Canada	Pierre L. Siklos
November	2000	The Relationship between the Federal Funds Rate and the Fed’s Funds Rate Target: Is it Open Market or Open Mouth Operations?	Daniel L. Thornton

November	2000	Expectations and the Stability Problem for Optimal Monetary Policies	George W. Evans Seppo Honkapohja
January	2001	Unemployment, Factor Substitution, and Capital Formation	Leo Kaas Leopold von Thadden
January	2001	Should the Individual Voting Records of Central Banks be Published?	Hans Gersbach Volker Hahn
January	2001	Voting Transparency and Conflicting Interests in Central Bank Councils	Hans Gersbach Volker Hahn
January	2001	Optimal Degrees of Transparency in Monetary Policymaking	Henrik Jensen
January	2001	Are Contemporary Central Banks Transparent about Economic Models and Objectives and What Difference Does it Make?	Alex Cukierman
February	2001	What can we learn about monetary policy transparency from financial market data?	Andrew Clare Roger Courtenay
March	2001	Budgetary Policy and Unemployment Dynamics	Leo Kaas Leopold von Thadden
March	2001	Investment Behaviour of German Equity Fund Managers – An Exploratory Analysis of Survey Data	Torsten Arnswald
April	2001	The information content of survey data on expected price developments for monetary policy	Christina Gerberding
May	2001	Exchange rate pass-through and real exchange rate in EU candidate countries	Zsolt Darvas

July	2001	Interbank lending and monetary policy Transmission: evidence for Germany	Michael Ehrmann Andreas Worms
September	2001	Precommitment, Transparency and Montetary Policy	Petra Geraats
September	2001	Ein disaggregierter Ansatz zur Berechnung konjunkturbereinigter Budgetsalden für Deutschland: Methoden und Ergebnisse *	Matthias Mohr
September	2001	Long-Run Links Among Money, Prices, and Output: World-Wide Evidence	Helmut Herwartz Hans-Eggert Reimers
November	2001	Currency Portfolios and Currency Exchange in a Search Economy	Ben Craig Christopher J. Waller
December	2001	The Financial System in the Czech Republic, Hungary and Poland after a Decade of Transition	Thomas Reininger Franz Schardax Martin Summer
December	2001	Monetary policy effects on bank loans in Germany: A panel-econometric analysis	Andreas Worms
December	2001	Financial systems and the role of banks in monetary policy transmission in the euro area	M. Ehrmann, L. Gambacorta J. Martinez-Pages P. Sevestre, A. Worms
December	2001	Monetary Transmission in Germany: New Perspectives on Financial Constraints and Investment Spending	Ulf von Kalckreuth
December	2001	Firm Investment and Monetary Transmission in the Euro Area	J.-B. Chatelain, A. Generale, I. Hernando, U. von Kalckreuth P. Vermeulen

***** Available in German only.

January	2002	Rent indices for housing in West Germany 1985 to 1998	Johannes Hoffmann Claudia Kurz
January	2002	Short-Term Capital, Economic Transformation, and EU Accession	Claudia M. Buch Lusine Lusinyan
January	2002	Fiscal Foundation of Convergence to European Union in Pre-Accession Transition Countries	László Halpern Judit Neményi
January	2002	Testing for Competition Among German Banks	Hannah S. Hempell
January	2002	The stable long-run CAPM and the cross-section of expected returns	Jeong-Ryeol Kim
February	2002	Pitfalls in the European Enlargement Process – Financial Instability and Real Divergence	Helmut Wagner
February	2002	The Empirical Performance of Option Based Densities of Foreign Exchange	Ben R. Craig Joachim G. Keller
February	2002	Evaluating Density Forecasts with an Application to Stock Market Returns	Gabriela de Raaij Burkhard Raunig

Visiting researcher at the Deutsche Bundesbank

The Deutsche Bundesbank in Frankfurt is looking for a visiting researcher. Visitors should prepare a research project during their stay at the Bundesbank. Candidates must hold a Ph D and be engaged in the field of either macroeconomics and monetary economics, financial markets or international economics. Proposed research projects should be from these fields. The visiting term will be from 3 to 6 months. Salary is commensurate with experience.

Applicants are requested to send a CV, copies of recent papers, letters of reference and a proposal for a research project to:

Deutsche Bundesbank
Personalabteilung
Wilhelm-Epstein-Str. 14

D - 60431 Frankfurt
GERMANY