Evaluating Risk Models with Likelihood Ratio Tests: Use with Care!

Gabriela de Raaij and Burkhard Raunig[*,**]

March, 2002

Gabriela de Raaij
Central Bank of Austria
Financial Markets Analysis Division
Otto-Wagner-Platz 3
POB 61, A-1011 Vienna
Austria
Phone: (+43-1) 404 20-3249
Fax: (+43-1) 404 20-3299
Email: gabriela.raaij@oenb.co.at

Burkhard Raunig
Central Bank of Austria
Economic Studies Division
Otto-Wagner-Platz 3
POB 61, A-1011 Vienna
Austria
Phone: (+43-1) 404 20-7219
Fax: (+43-1) 404 20-7299
Email: burkhard.raunig@oenb.co.at

*) Corresponding author

**) The opinions expressed do not necessarily reflect those of the Austrian Central Bank.

Evaluating Risk Models with Likelihood Ratio Tests: Use with Care!

Abstract

Most modern approaches to measure and control the risks of financial portfolios are either directly or indirectly based on density forecasts. Tools to evaluate the quality of such forecasts are therefore essential. In this paper we examine a recently proposed methodology to evaluate density forecasts from risk models that builds on likelihood ratio tests. We discuss three cases that are highly relevant in risk management where likelihood ratio tests fail to detect incorrect density forecasts. We illustrate this fact with Monte Carlo simulations and empirical examples. We also demonstrate that the likelihood ratio testing framework in conjunction with additional diagnostic tests is an attractive tool to evaluate risk models.

1 Introduction

Traditionally, the forecast evaluation literature has primarily dealt with methods to evaluate point forecasts. However, over the last few years interest by the financial industry has increased into density forecasts. Financial institutions became interested to supplement standard risk measures as for example portfolio variance and correlation with broader information on portfolio risk. Especially in the area of risk management density forecasts are frequently generated since they provide a full picture of the uncertainty associated with a portfolio. Therefore, density forecasts and measures derived from such forecasts play a key role in modern risk management. In particular, Value at Risk (VaR), which is defined as a certain quantile of a forecast of the entire return distribution of a financial portfolio (1% and 5% quantiles are typically used) has become the backbone of modern risk management (Jorion, 1996, Duffie and Pan, 1997). Moreover, regulatory authorities have permitted banks to use VaR estimates to determine their capital requirements to cover their exposure to market risk. Therefore, perhaps not surprisingly, techniques to evaluate the quality of such forecasts are of paramount importance for internal as well as regulatory purposes.

Various methods to evaluate density forecasts have been proposed in the literature. Methods that evaluate Value at Risk estimates directly have been proposed and examined in Kupiec (1995), Lopez (1998), Christoffersen (1998) and Christoffersen, Hahn and Inoue (2001). More general evaluation methodologies that take a broader view and consider the whole distribution instead of just a single quantile have recently been proposed in Crnkovic and Drachman (1997) and Diebold, Gunther and Tay (1998). In this paper we focus on the second kind of methodologies that evaluate density forecasts via the entire forecasted distribution. We examine an interesting extension of Diebold et all. developed in Berkowitz (2001) that suggests statistical tests of the quality of density forecasts within a likelihood ratio (LR) framework.

Although the LR-framework is attractive, there are important cases where the uncritical use of this framework or equivalent test procedures may lead to erroneous conclusions about the quality of density forecasts. We outline three cases where deficient density forecasts cannot be detected within the LR-framework and relate them to the evaluation of VaR models. In these cases variance/covariance models and historical simulation models to estimate VaR may not be rejected even if they deliver poor density forecasts. Using Monte Carlo simulations and an empirical illustration we highlight that in the three cases the basic LR-framework alone as well as an extended LR test that covers higher order dependencies and certain kinds of nonlinearities has little power to detect incorrect density forecasts. However, we also demonstrate that the LR framework in conjunction with additional diagnostic tests is a constructive and powerful framework to identify deficient forecasting models.

The rest of the paper is organized as follows. Section 2 outlines the LR density forecast evaluation framework of Berkowitz (2001). The three cases that we consider are discussed in section 3. The Monte Carlo experiments and the empirical examples are reported in section 4. Some final remarks are provided in section 5.

2 Density Forecast Evaluation and the LR Framework

Let $\{x_t\}_{t = 1,..., m}$ be a time series generated from the 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 $\{x_t\}_{t = 1,..., m}$.[1] The quality of such forecasts can be evaluated with the help of a probability integral transformation (PIT) suggested in 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] 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.

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

Diebold, Gunther and Tay (1998) show that the transformed series $\{z_t\}_{t = 1,...,m}$ must be independently and identically uniformly distributed (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 conditional densities $\{f_t(x_t)\}_{t = 1,..., m}$.[2]

Hence, the quality of density forecasts can be assessed by an examination of the properties of the z-series resulting from the PIT given by equation (1). Such examinations can either be based on descriptive diagnostic tools or on statistical tests as proposed in Crnkovic and Drachman (1997). Diebold et al. advocate graphical methods. However, there may be situations in which statistical testing is required. For example, within a financial institution one may have to compare the quality of Value at Risk forecasts across different trading books with the help of formal test procedures. Another example may be a regulatory authority that wants to assess the accuracy of risk measurement systems of different financial institutions. To assure a uniform treatment across the involved institutions the authority may therefore ask them to carry out statistical tests for a portfolio of financial instruments as defined by the supervision authority.

Berkowitz (2000) emphasizes that statistical tests that are directly based on a z-series require rather large sample sizes to be reliable and suggests a further transformation of the individual z_t's to obtain more powerful test statistics. The transformation for a single z_t is given by

$$n_t = \Phi^{-1}(z_t), \qquad (2)$$

where $\Phi^{-1}(.)$ denotes the inverse of a standard normal distribution function. This transformation produces an n-series that is independently standard normally distributed (iid

[2] 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).

N(0,1)) if the true- and the forecasted conditional distributions coincide. Berkowitz proposes likelihood-ratio tests against the first order autoregressive alternative

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

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

$$LR = -2\left(L(0,1,0) - L\left(\hat{\mu}, \hat{\sigma}^2, \hat{\rho}\right)\right) \sim \chi^2(3), \tag{4}$$

where σ^2 is the variance of ε_t and L(.) denotes a Gaussian log-likelihood function. In simulation experiments he demonstrates that the test statistic has good small sample properties.

He also suggests an extended LR test that covers the possibility of higher order dependence as well as nonlinear dependencies that may be constructed from the model

$$n_t = \alpha_0 + \alpha_1 n_{t-1} + ... + \alpha_m n_{t-m} + \beta_1 n_{t-1}^2 + ... + \beta_h n_{t-h}^2 + \varepsilon_t . \tag{5}$$

Acceptance of the hypotheses $\alpha_0 = ... = \alpha_m = \beta_1 = ... = \beta_h = 0, \mathrm{Var}(\varepsilon_t) = 1$ would indicate correct density forecasts. Specifications of this type could easily be extended to include more lags, higher powers of lagged n_t's or cross products of lagged n_t's. Individual or joint hypotheses about models (4) or (5) could also be tested within a regression framework using standard t-, F- and chi square tests.

3 Three Critical Cases

The LR-tests based on equations (4) or (5) or equivalent test procedures are attractive because they are easy to implement. Especially tests based on a setting like equation (5) appear to be quite general. However, the evidence from such tests must be interpreted with care. We now discuss three cases where the uncritical use of the LR-tests described above may lead to erroneous conclusions about the quality of density forecasts. We also point out how additional diagnostic tests may help to identify misspecified forecasting models. Since the discussion is

not only relevant for the evaluation of risk models but also important for the evaluation of the forecasting ability of time series models in general, we first describe the cases in a time series setting. Hereafter we point out how these cases may arise in the evaluation of a risk measurement system.

We start by assuming that a time series of financial returns is generated by the (possibly) nonlinear time series model

$$x_t = \mu(I_{t-1}) + \sigma(I_{t-1})\xi_t, \tag{6}$$

where $\mu(.)$ is the conditional mean and $\sigma(.)$ is the conditional standard deviation. The available information set is denoted by I_{t-1} and the error term ξ_t is assumed to be a conditionally standardized martingale difference sequence (i.e. $E(\xi_t|I_{t-1}) = 0$ and $Var(\xi_t|I_{t-1}) = 1$). As discussed in Bai and Ng (2001) this framework encompasses most standard linear and nonlinear time series models with or without exogeneous explanatory variables, traditional ARMA models and models with ARCH and GARCH disturbances. This model can also be interpreted as a representation for the evolution of the returns of a portfolio of financial instruments over time. Let us now describe the different cases within this framework. We assume that model (6) is the true data generating process.

Case 1: Suppose that a density forecaster has generated density forecasts from a misspecified version of model (6). He has correctly specified the first and second conditional moments. However, he incorrectly assumes that the error terms ξ_t^* are N(0,1) whereas the true error terms ξ_t are uncorrelated with mean zero and unit standard deviation but have a (possible time varying) distribution $D_t(0,1) \neq N(0,1)$ which differs from a normal distribution for all t. To assess the quality of his forecasts he applies the transformations given by (1) and (2) and performs an LR test of the kind described in section 2. Since the transformations simply produce $n_t = \Phi^{-1}(\Phi(\xi_t^*)) = \xi_t^*$ he will obtain an n-series that has a zero conditional mean, is

uncorrelated and has a unit standard deviation.[3] Because the LR test maintains the assumption of normality, the test will not reject although the n-series is not normally distributed and the density forecasts are incorrect. Without additional distributional tests for normality of an n-series the inadequate density forecasts will therefore erroneously be accepted despite the fact that the true densities are definitely not normal.

Case 2: Now assume that a density forecaster has issued density forecasts that only capture the conditional mean correctly. He erroneously assumes normally distributed density forecasts with constant unconditional standard deviation σ. Transformations (1) and (2) produce an n-series that is uncorrelated with conditional and unconditional zero mean and a unit unconditional standard deviation if the estimated unconditional standard deviation σ is correct. However, the resulting n-series is heteroskedastic if the true conditional variance $\sigma(I_{t-1})$ is time varying.[4] Since LR tests based on (4) or (5) are not designed to detect heteroskedasticity, the forecaster will not be able to identify the inability of the forecasts to capture the true volatility dynamics without additional tests for heteroskedasticity. Of course, the neglected volatility dynamics of the density forecasts is likely to be reflected in the shape of the unconditional distribution of the resulting n-series because the time varying volatility tends to produce n-series with a fat tailed distribution. However, the presence of a fat-tailed distribution alone does not help to discriminate between an incorrect volatility dynamics and an incorrect shape of the conditional distributions.

Case 3: Finally, assume that the forecaster correctly specifies the mean dynamics but uses the unconditional density of $\{x_t\}_{t = 1,\ldots,m}$ instead of the conditional densities as a density forecasts for future x_t's. If the true process is stationary then the integral transformation (1) with respect

[3] For a similar result in the context of a z-series, see Diebold, Hahn and Tay (1999).
[4] This can be shown by noting that $s_t = [x_t - \mu(I_{t-1})]/\sigma$ and $n_t = \Phi^{-1}(\Phi(s_t)) = s_t$. Using these relationships it can be shown that $E(n_t^2|I_{t-1}) = \sigma(I_{t-1})/\sigma$, $E(n_t|I_{t-1}) = 0$, $E(n_t\, n_{t-j}) = 0$, $E(n_t) = 0$ and $E(n_t^2) = 1$.

to the unconditional distribution will produce an uncorrelated and almost perfectly uniformly distributed z-series.[5] The subsequent transformation (2) with the inverse of the standard normal distribution will then of course create an uncorrelated n-series that has a distribution quite close to a standard normal distribution. The distribution of the n-series will be virtually standard normal despite the fact that the volatility dynamics is misspecified because the unconditional distribution of the original data ignores the order in which the observations are arranged. Given such a situation, neither the LR tests nor additional distributional tests for normality will indicate incorrect density forecasts. One way to detect an incorrect volatility dynamics of the density forecasts is to examine the time series of squared n_t's which will display clustering if the volatility dynamics has been neglected. The n-series will also be standard normal if the true conditional densities change over time due to other time dependent higher moments because this is also already reflected in the shape of the unconditional distribution. To identify additional deficiencies higher powers of the n-series would have to be investigated.

Having outlined the three cases in a time series setting we now point out how these cases may arise in an evaluation of the quality of a VaR-model. Consider a financial institution that uses a variance/covariance-model to calculate its daily Value at Risk (for a comprehensive discussion of this approach, see Jorion, 1997). In this risk model the VaR of a financial portfolio is a certain multiple of the forecasted conditional standard deviation of the distribution of the portfolio returns. It is typically assumed that the returns of the portfolio follow a conditional normal distribution. For correct VaR calculations the correct estimation of the portfolio volatility and the correctness of the normal distribution assumption are critical. Now suppose that the VaR model adequately captures the volatility dynamics of the

[5] Because the unconditional distribution can only be estimated (for example with the empirical distribution function), small deviations from the uniform distribution may result from estimation errors. These errors are likely to be small if a sufficiently large number of observations are available.

portfolio but the normal distribution assumption for the portfolio does not hold.[6] This may happen if the portfolio contains a significant amount of nonlinear instruments such as options or if the underlying risk factors are already not normally distributed. This situation obviously corresponds to case 1 outlined above. Therefore, it is very likely that the LR tests or equivalent tests do not to reject and give the impression that the VaR model is adequate despite the fact that the true VaR may be far away from the VaR estimated with the model.

An even more drastic example that corresponds with case 2 is a quite naïve variance/covariance model which is again built on the assumption of normally distributed portfolio returns and it is further assumed that the portfolio variance is simply constant. If the unconditional portfolio variance is estimated correctly then the LR tests will not reject the VaR model even if both assumptions are clearly violated.

Case 3 may arise if a financial institution uses a historical simulation for the purpose of VaR calculations. In the historical simulation the past observations of a set of risk factors are interpreted as possible future realizations of the risk factors. The return on a portfolio of financial instruments is computed under each of the historical scenarios of the risk factors and the VaR is then calculated as a certain quantile of the resulting portfolio return distribution. To obtain accurate VaR estimates 500, 1000 or even more historical realizations are often used. Since each historical scenario is equally weighted the resulting VaR is implicitly based on an estimate of the unconditional return distribution of the portfolio (Hull and White, 1998 and Huisman, Koedijk and Pownal, 1998). If the unconditional distribution is stationary or only very slowly changing then for a fixed portfolio the marginal distribution of the resulting n-series used in the LR tests will be close to a standard normal distribution even if volatility is time varying. The LR tests will again not reject the null hypothesis of a correct risk model despite the fact that the volatility dynamics is ignored by the model.

[6] For simplicity we assume in the discussion that the portfolio returns have zero mean.

4 Simulations and Empirical Illustrations

We illustrate the three cases where the LR tests fail to identify incorrect density forecasts with simulation experiments and empirical examples for the daily returns on the S&P 500 and the FTSE 30 stock market indices. In the Monte Carlo simulations we consider a GARCH(1,1)-t model

$$r_t = \sqrt{\sigma_t^2}[\nu/(\nu-2)]^{-1/2}\varepsilon_t \qquad \varepsilon_t \sim t_5$$

$$\sigma_t^2 = 0.004 + 0.03\varepsilon_{t-1}^2 + 0.95\,\sigma_{t-1}^2,$$

where r_t denotes the simulated returns, σ_t denotes the conditional standard deviation and ε_t denotes an innovation drawn from a student t distribution with $\nu = 5$ degrees of freedom. This model is a standard model for financial returns. It implies fat tailed student t distributed density forecasts and produces the volatility clustering often observed in financial return series.

We use the simulated time series from the model to examine the three cases outlined in section 3. In case 1 we assume conditionally normally distributed density forecasts instead of student t_5 distributed forecasts, and the GARCH(1,1) model is estimated under this incorrect distributional assumption. Since the estimated model parameters are still consistent (Bollerslev and Woooldrige, 1992, Lumsdaine, 1996) the volatility dynamics should be adequately captured despite the fact that the true distribution is a fat tailed student t distribution. In case 2 we incorrectly assume an unconditional normal distribution for the density forecasts. We thereby, in addition to the choice of an incorrect distribution, also misspecify the volatility dynamics because we use a simple estimate of the unconditional standard deviation instead of an estimate of the time dependent conditional variance. In the third case we take the empirical distribution function as our density forecasts and therefore again misspecify the conditional distributions of our density forecasts.

We perform 10000 simulations of the model for samples of 500, 1000, 2000 and 4000 observations. For each of the three cases we estimate the rejection rates of a likelihood ratio test (LR1) based on (4) and a likelihood ratio test (LR2) based on two lags of n_t and n_t^2 that considers the more general alternative (5) when applied to the resulting n-series for a 5% significance level. Using the same significance level, we also calculate the rejection rates of a Jarque-Bera normality test and an ARCH test for heteroskedasticity (ARCH) based on an F test of the restriction $\gamma_1 = \gamma_2 = \ldots = \gamma_5 = 0$ in the regression $n^2_t = \gamma_0 + \gamma_1 n^2_{t-1} + \ldots + n^2_{t-5} + \xi_t$. The results of the simulation experiments are reported in table 1.

INSERT TABLE 1 ABOUT HERE

From table 1 it is easily seen that in all three cases the LR1 and LR2 tests have little power to detect incorrect density forecasts. This is exactly what we would expect from our discussion in section 3. For example, the rejection rates of the LR1 test are extremely low and range from 0.02 to 0.041 across the different sample sizes and cases. The rejection rates of the more general LR2 test are similar to the LR1 rejection rates in case 1 and slightly higher, but still very low, in the other two cases. On the other hand, note that the JB test virtually always rejects the incorrect density forecasts in case 1 and case 2 and never rejects in case 3. This finding is again consistent with the theoretical discussion. The same is true for the heteroskedasticity tests. The ARCH test virtually never rejects in case 1 because the volatility dynamics is correctly captured by the forecasts but rejects frequently in the other two cases where density forecasts ignore the volatility dynamics. Taken together, the results from the simulations clearly show that both LR tests are essentially unable to identify the incorrect density forecasts in each case. Without the additional normality- and heteroskedasticity tests the deficient forecasts cannot be detected.

Let us now turn to the empirical example. We evaluate successive one-step-ahead density forecasts for daily returns on the FTSE 30 and the S&P 500 from four different models. The first two models are the simple moving average model (MA) of squared returns

with a rolling time window of 250 trading days and the exponentially weighted moving average model of squared returns (EWMA) with a decay factor of 0.94, as suggested by J. P. Morgan. In both models we make the conventional assumptions that the mean of the daily returns is approximately zero and that the returns are conditionally normally distributed. These models are often used to generate the variance/covariance matrices used in VaR calculations. The other two models are the standard GARCH(1,1)-n model where the errors are also assumed to be conditionally normal and the GARCH(1,1)-t model where conditionally t distributed errors are assumed.

INSERT TABLE 2 ABOUT HERE!

The likelihood ratio-, normality- and heteroskedasticity tests for the n-series resulting from the 1,000 daily density forecasts of the different models over the period from 4/20/1998 to 2/15/2002 for the S&P 500 and 4/21/1998 to 2/18/2002 for the FTSE 30 are summarized in table 2. The empirical evidence again highlights the inability of the LR tests to discriminate between the density forecasts from the different models. For example, the LR1 and LR2 tests do not distinguish between the GARCH-n and the GARCH-t models.[7] The tests indicate correct density forecasts for both models. However, the additional JB- and ARCH tests suggest that only the density forecasts from the GARCH-t models might be correct. The forecasts of the GARCH-n model are clearly rejected by the JB test. Since the ARCH tests do not reject for the GARCH-n model the normality assumption appears to be incorrect. In the case of the S&P 500 the LR1 test does also not reject the simple MA model although the JB- and ARCH tests clearly indicate that both, the volatility dynamics and the normal distribution assumption are incorrect. The empirical evidence in general suggests that the widely used MA- and EWMA models combined with the assumption of a normal distribution do not produce accurate density forecasts.

[7] The parameter estimates for the GARCH models are available on request from the authors.

Concluding Remarks

In this paper we investigated a recently proposed likelihood ratio framework to evaluate density forecasts. We showed that the uncritical use of this framework may lead to incorrect conclusions about the quality of risk models. Standard variance/covariance approaches and historical simulation approaches to calculate VaR may be accepted despite the fact that they may provide poor VaR estimates. We further demonstrated that additional diagnostic tests including normality tests and tests for heteroskedasticity help to detect incorrect models. But these additional tests do not only help to detect incorrect models, they also provide information about the kind of model failure. This leads us to the conclusion that the LR framework of Berkowitz combined with additional diagnostic tests is a constructive and powerful tool to evaluate risk models. Of course, a careful risk manager would probably also perform some of the other tests mentioned in the introduction to asses the accuracy of his risk model. However, if one of the cases outlined in this paper arises, he may obtain conflicting results if he compares the outcome of these tests with the evidence from the LR tests. With the help of further graphical assessments or the additional diagnostic tests he may then be able to correctly interpret the results.

Bai, J. & Ng, S. (2001). A Consistent Test for Conditional Symetry in Time Series Models. Journal of Econometrics, 103, 225-258.

Berkowitz, J. (2001). Testing Density Forecasts, with Applications to Risk Management. Journal of Business and Economic Statistics, 19(4), 465-475.

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.

Christoffersen, P. (1998). Evaluating Interval Forecasts. International Economic Review, 841-862.

Christoffersen, P., Hahn J. & Inoue A. (2001). Testing and Comparing Value at Risk Measures. Journal of Empirical Finance, 8 (July), 325-342.

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., 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.

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.

Duffie, D. & Pan, J. (1997). An Overview of Value at Risk. Journal of Derivatives 4 (Spring), 7-49.

Huisman, R., Koedijk, K. G. & Pownal, R. A. J. (1998). VaR+: Fat Tails in Financial Risk Management. Journal of Risk, 1 (Fall), 47-61.

Hull, J. & White, A. (1998). Incorporating Volatility Updating Into the Historical Simulation Method for Value at Risk, Journal of Risk,1 (Fall), 1-19.

Jorion, P. (1996). Value at Risk: The New Benchmark for Controling Risk. Irwin Professional.

Kupiec, P. (1995). Techniques for Verifying the Accuracy of Risk Measurement Models. Journal of Derivatives 3 (Winter), 73-84.

Lopez, J. (1996). Regulatory Evaluation of Value at Risk Models. FRBNY Economic Policy Review 4 (October), 119-124.

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.

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

Table 1: Rejection Rates of Density Forecasts from GARCH(1,1)-t Model, 5% Significance Level

	Observations	**LR1**	**LR2**	**JB**	**ARCH**
Case 1	4000	0.020	0.027	1.000	0.045
	2000	0.020	0.031	1.000	0.042
	1000	0.022	0.029	1.000	0.040
	500	0.026	0.028	0.990	0.039
Case 2	4000	0.041	0.164	1.000	0.930
	2000	0.034	0.131	1.000	0.706
	1000	0.032	0.107	1.000	0.440
	500	0.030	0.083	0.993	0.257
Case 3	4000	0.025	0.081	0.000	0.993
	2000	0.027	0.084	0.000	0.873
	1000	0.026	0.081	0.000	0.581
	500	0.024	0.073	0.000	0.317

Notes: The table reports the rejection rates of density forecast tests from 10,000 simulations of the model $r_t = \sqrt{\sigma_t}\varepsilon_t$, $\sigma_t^2 = 0.004 + 0.03\varepsilon^2_{t-1} + 0.95\ \sigma^2_{t-1}$ where the innovations ε_t are drawn from a student t distribution with 5 degrees of freedom. LR1 and LR2 denote the likelihood ratio tests based on equations (4) and (5) in the text, respectively. JB denotes a Jarque-Bera test for a normal distribution. ARCH denotes an F test for heteroskedasticity of the restriction $\gamma_1 = \gamma_2 = \ldots = \gamma_5 = 0$ in the regression $n^2_t = \gamma_0 + \gamma_1 n^2_{t-1} + \ldots + n^2_{t-5} + \xi_t$.

Table 2: P-Values from Evaluations of Density Forecasts from MA-, EWMA-, GARCH(1,1)-n and GARCH(1,1)-t Models for Daily Returns on the FTSE 30 and the S&P 500.

FTSE 30 Period: 4/21/1998 to 2/18/2002 Observations: 1000

Model	**LR1**	**LR2**	**JB**	**ARCH**
MA	0.000	0.000	0.000	0.000
EWMA	0.000	0.003	0.000	0.618
GARCH-n	0.865	0.625	0.000	0.374
GARCH-t	0.964	0.507	0.221	0.584

S&P 500 Period: 4/20/1998 to 2/15/2002 Observations: 1000

Model	**LR1**	**LR2**	**JB**	**ARCH**
MA	0.754	0.046	0.000	0.000
EWMA	0.063	0.060	0.000	0.402
GARCH-n	0.759	0.431	0.000	0.219
GARCH-t	0.873	0.626	0.203	0.256

Notes: The table reports p-values from tests of the quality of 1000 consecutive one step ahead density forecasts from a moving average volatility model (MA) with a rolling window of 250 trading days, an exponentially weighted moving average volatility model (EWMA) with decay factor 0.94, a GARCH(1,1) model with normally distributed errors (GARCH-n) and a GARCH(1,1) model with t-distributed errors (GARCH-t). LR1 and LR2 denote the likelihood ratio tests based on equations (4) and (5) in the text, respectively. JB denotes a Jarque-Bera test for a normal distribution. ARCH denotes an F test for heteroskedasticity of the restriction $\gamma_1 = \gamma_2 = \ldots = \gamma_5 = 0$ in the regression $n^2_t = \gamma_0 + \gamma_1 n^2_{t-1} + \ldots + n^2_{t-5} + \xi_t$.