Time Series: Forecasting and Evaluation Methods With Concentration On Evaluation Methods for Density Forecasting

Master's thesis in Statistics

Financial Theory and Insurance Mathematics

Therese Grindheim

Supervisor

Yushu Li

Department of Mathematics

University of Bergen

May 2018

Abstract

The main focus of this thesis are density forecasts and the corresponding evaluation methods. A density forecast is an estimate of the probability density of predicted values. Density forecasts and the related evaluation methods have been little explored compared to point and interval forecasts, therefore we have chosen to focus on this topic. We go through a detailed description of three evaluation methods for density forecasts. To measure the performance of two of the density forecast evaluation methods we perform a Monte Carlo simulation. We simulate data sets with different data generating mechanisms to measure the size and power for the chosen evaluation methods. Based on our results from the Monte Carlo simulation, we continue with one evaluation method and apply it on empirical data, more specifically on economical, financial and insurance time series data.

Acknowledgments

First, I want to thank my supervisor Yushu Li for great guidance and assistance throughout the course of writing this thesis. Also, many thanks to my family and partner for the support and patience during this process. And finally, a great thank you to my fellow students, Pernille Kiil, Victoria Foster and Helen Bringeland, for keeping my spirits up during our many lunch breaks.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This paper will focus on methods for forecasting and evaluation methods for forecasts in the framework of time series data. In Chapter 2, we will start by giving a short overview of important concepts of time series and forecasting, continued with an introduction to three different types of forecasts: point, interval and density forecasts. In the introduction we discuss shortly the advantages and shortfalls of the different types of forecasting.

One of the main objectives when discussing time series analysis is forecasting. Forecasting, in terms of time series, is trying to estimate the future values of the series based on the available observations (Chatfield, 2000). A large amount of research has focused on producing and evaluating point forecasts, for example Fuller and Hasza (1981), Engle and Yoo (1987) and Gneiting (2001). Point forecasts only provides the most likely outcome for the predicted variable and doesn't give any information around the uncertainty around the predicted outcome (Montgomery, Jennings and Kulahci, 2016). Point forecasts are often the first-order importance for a forecast user since they are not very difficult to produce compared to interval and density forecasts and are easy to understand (Christofferesen, 1998). Since a forecast user can't be sure about the predicted outcome until it is observed and might want

to plan different strategies based on different outcomes, a point forecasts can be inadequate. A first reaction to the lack of description of certainty from the point forecasts are the interval forecasts (Tay and Wallis, 2000). An interval forecast can produce the most likely range of outcomes at different confidence intervals and therefore the forecast users can plan different strategies for different outcomes. In Chatfield's (1993) article we find a detailed description of methods for a forecast user to produce interval forecasts based on different assumptions. An even more descriptive forecast is the density forecast. The density forecast that can provide a complete description of the probabilities for the different outcomes of a predicted variable and the interest in these forecasts has increased in the recent years (Tay and Wallis, 2000). There are different possibilities for a forecast user to produce density forecasts, e.g. based on assumptions on the error term, based on simulations and based on surveys.

Since evaluation of forecasts are of great importance when discussing forecast, this paper will focus on different evaluation methods corresponding to the different forecasts. A lot of literature has focused on how to produce and evaluate point forecasts, therefore we will focus more on methods for interval and density forecasts. Especially density forecasting has been little explored compared to point and interval forecasting.

In Chapter 3 we will go through evaluation methods for point and interval forecasts. The original method for evaluating interval forecasts only evaluate the unconditional coverage. For time series interval forecasts it is often inadequate to only consider the unconditional coverage, therefore Christofferesen presented a method for evaluating interval forecasts testing for unconditional coverage, independence and conditional coverage (Wallis, 2003). The test for conditional coverage uses a likelihood ratio framework and is a combination of the unconditional coverage and the independence tests.

We will go through the details of three different methods for evaluating density forecasts

in Chapter 4. The first method we go through was presented by Diebold, Gunther and Tay (1998). The method is based on the probability integral transformation and uniform distribution. The test is for unconditional coverage for the distribution and it can be supplemented with visual tools to investigate independence and coverage by the uniform distribution for the transformed variables. Berkowitz (2001) recognized that it is difficult to test for uniformity in small data samples and introduced a method for evaluation density forecasts based on transformation to the normal distribution. His method consists of two tests, one for conditional coverage and one for independence. The last method we go through was introduced by Li and Andersson (2018) and it is an extension of Christoffersen (1998) method for evaluation of interval forecasts. It consists of three tests, one for unconditional coverage, one for independence and one for conditional coverage. The test for conditional coverage is a combination of the independence test and the unconditional coverage test. The method does not need a parametric specification for the time dynamics.

To see how well various density forecast evaluation methods work we will perform a Monte Carlo study and present the results in Chapter 5. We will simulate data sets with different data generating mechanisms, and measure the size and power of the evaluation methods for density forecasting. We will compare Berkowitz's (2001) method with Li and Andersson's (2018) method to see how well they perform. Since Li and Andersson (2018) also proposed a test for unconditional coverage, we will compare this test with the Kolmogorov-Smirnov test (Chkravarti, Laha and Rot, 1967).

In Chapter 6, we will use one of the methods for density forecasting to evaluate economical, financial and insurance data. We will continue with the tests with the best results from the Monte Carlo study when evaluating the data sets. We have four data sets we will evaluate in the empirical studies. We have three financial and economical data sets. The three sets consist of a data set of a quarterly time series for the real gross domestic product in the US, a data set

of a daily time series of the Standard&Poor500 Index and a data set of the transformed log-returns for a daily time series of the New York Stock Exchange Amex Composite Index. The fourth set is an insurance time series of compensations amount fire damage incidents caused by electronic equipment. We will test for the hypotheses of conditional and unconditional coverage with specified distributions and independence for the four data sets. The final chapter is a summary of the paper and the results obtained in the different studies.

Chapter 2

Time Series and Forecasting

2.1 Time Series

2.1.1 Introduction to Time Series

This paper will focus on forecasting with time series and applications in economics, finance and insurance. We start with an introduction to the main concepts and properties of time series and some important time series models. A time series is a sequence of stochastic variables, $\{Y_t\}_{t=1}^{T}$, listed in time order. A time series sequence is usually taken at regular time intervals. There is not a minimum or maximum amount of time that must be included in the sequence, so the user can gather the data points in a way that provides enough information for the analysis. Time series often occur naturally in economics, finance and insurance, for example quarterly data for unemployment, daily exchange rate and monthly rate of traffic accidents that result in insurance claims etc.

Following is a brief introduction to the main properties of a time series. The following definitions are from Brockwell and Davis (2016), and these concepts will be used in later

chapters in this thesis:

Definition 2.1.1. *Moments of a Time Series*

Let $\{Y_t\}_{t=1}^T$ be a time series with $E(Y_t^2) < \infty$, then the mean, variance, covariance and correlation functions of $\{Y_t\}$ are defined as :

$$\mu_t = E(Y_t) \tag{2.1.1}$$

$$\sigma_t^2 = E[(Y_t - \mu_t)^2] \tag{2.1.2}$$

$$\gamma_t(j) = E[(Y_t - \mu_t)(Y_{t+j} - \mu_{t+j})] \tag{2.1.3}$$

$$\rho_t(j) = \frac{\gamma_t(j)}{\sigma_t \sigma_{t+j}}. \tag{2.1.4}$$

Definition 2.1.2. *Lag Operator*

The lag or backward shift operator is a linear operator and has the following properties for a time series $\{Y_t\}_{t=1}^T$:

$$LY_t = Y_{t-1} \tag{2.1.5}$$

$$L^{-1}Y_t = Y_{t+1} \tag{2.1.6}$$

$$L^k Y_t = Y_{t-k}. \tag{2.1.7}$$

Definition 2.1.3. *Difference Operator*

Let $\{Y_t\}_{t=1}^T$ be a time series, then the first, second and d difference operators for $\{Y_t\}$ are defined as:

$$\Delta Y_t = Y_t - Y_{t-1} = (1 - L)Y_t \tag{2.1.8}$$

$$\Delta^2 Y_t = Y_t - 2Y_{t-1} + Y_{t-2} = (1 - 2L + L^2)Y_t = (1 - L)^2 Y_t \tag{2.1.9}$$

$$\Delta^d Y_t = (1 - L)^d Y_t, \ d > 0. \tag{2.1.10}$$

Definition 2.1.4. *Strict Stationarity*

A time series can have many different properties and one of them is that the process can be stationary. Generally speaking a time series can be viewed as stationary if there is

no systematic change in variance, no systematic trend and seasonality doesn't exist. If the series $\{Y_t\}_{t=1}^{T}$ is strict stationary, then the joint distribution of $Y(t_1), \ldots, Y(t_n)$ and the joint distribution of $Y(t_{1+h}), \ldots, Y(t_{n+h})$ should be equal $\forall\ t_1, \ldots, t_n$ and h.

Definition 2.1.5. *Weak Stationarity*

A less restrictive stationary property is covariance stationarity, also called weak stationarity. A time series $\{Y_t\}_{t=1}^{T}$ is weakly stationary if it has the following properties:

$$E(Y(t)) = \mu\ \forall\ t \tag{2.1.11}$$

$$cov(Y(t), Y(t+h)) = \gamma(h)\ \forall\ t \tag{2.1.12}$$

i.e. the first two moments are independent of time t and do not depend on the position of Y_t.

Example 2.1.1. *White Noise*

One example of a stationary time series is the white noise process. The white noise process $\{\epsilon_t\}$, has the following properties:

$$E(\epsilon_t) = 0\ \forall\ t \tag{2.1.13}$$

$$Var(\epsilon_t) = \sigma^2\ \forall\ t \tag{2.1.14}$$

$$Cov(\epsilon_t, \epsilon_s) = 0\ if\ t \neq s \tag{2.1.15}$$

i.e. the process has mean zero, constant variance and is uncorrelated.

2.1.2 Time Series Models

We can build different time series models to capture the structure of the time series data, and one of the more common models is the autoregressive moving average ($ARMA$) model. The $ARMA$ model is a combination of the autoregressive (AR) model and the moving average (MA) model. The models can be used to model the mean, the first order moment, of a time

series. $ARMA$ models are used to model processes that are stationary. When a process is non-stationary, which is often the case with economical and financial data, one option is to model the data with the autoregressive integrated moving average ($ARIMA$) model. To model the standard deviation of a process we can use the autoregressive conditional heteroskedasticity ($ARCH$) model or the generalized autoregressive conditional heteroskedasticity ($GARCH$) model. In later chapters, when we are simulating data for the Monte Carlo simulation and in the studies of the empirical data sets, we will apply some of these common, popular time series models.

The following definitions for time series models are from Wei (1990).

Definition 2.1.6. *Autoregressive Model*

The autoregressive model with order p, $AR(p)$, for a time series $\{Y_t\}_{t=1}^{T}$ is defined as

$$Y_t = c + \sum_{i=1}^{p} \phi_i Y_{t-i} + \epsilon_t \tag{2.1.16}$$

where $\{\epsilon_t\}$ is a white noise process. We can check if the data is stationary by factoring

$$(1 - \phi_1 L - \cdots - \phi_q L^q) = (1 - \lambda_1 L) \ldots (1 - \lambda_q L) \tag{2.1.17}$$

and it is stationary if $|\lambda_1| < 1, \ldots, |\lambda_q| < 1$. Another way to check for stationarity is to see if all the characteristic roots of the lag operators

$$(1 - \phi_1 z - \cdots - \phi_q z^q) = 0 \tag{2.1.18}$$

lie outside the unit circle. The AR model can be used to model time series where the outcomes in present value depends on the past values and a random shock.

Definition 2.1.7. *Moving Average Model*

The moving average model with order q, $MA(q)$, for a time series $\{Y_t\}_{t=1}^{T}$ is defined as

$$Y_t = \mu + \epsilon_t + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} \tag{2.1.19}$$

where $\{\epsilon_t\}$ is a white noise process. One property of the $MA(q)$ model is that it can be invertible. If the model is invertible the $MA(q)$ can be written as an $AR(\infty)$. To check if the $MA(q)$ model is invertible, we set the mean to be zero, $\mu = 0$, then $Y_t = (1+\theta_1 L+\cdots+\theta_q L^q)\epsilon_t$. If we factorize the model

$$(1 + \theta_1 L + \cdots + \theta_q L^q) = (1 - \lambda_1 L) \ldots (1 - \lambda_q L) \tag{2.1.20}$$

and $|\lambda_1| < 1, \ldots, |\lambda_q| < 1$ the model is invertible. We can also check this property by controlling that the roots of the lag polynomial

$$(1 + \theta_1 z + \cdots + \theta_q z^q) = 0 \tag{2.1.21}$$

lie outside the unit circle. The MA model can be used to model time series where random events have an immediate impact and the impact has a short lived effect.

Definition 2.1.8. *Autoregressive Moving Average Model*

The $ARMA(p, q)$ model of order p and q is a combination of the $AR(p)$ and $MA(q)$ models and it is defined as for a time series $\{Y_t\}_{t=1}^T$

$$Y_t = \mu + \sum_{i=1}^{p} \phi_i Y_{t-i} + \sum_{j=0}^{q} \theta_j \epsilon_{t-j} \tag{2.1.22}$$

where $\{\epsilon_t\}$ is a white noise process. The $ARMA$ model is stationary and invertible for certain values of ϕ and θ. The stationary part of the $ARMA(p, q)$ model depends on the autoregressive part and if the roots of

$$1 - \phi_1 z - \cdots - \phi_p z^p = 0 \tag{2.1.23}$$

are all outside the unit circle the model is stationary. For the invertibility we require that the roots of

$$1 + \theta_1 z + \cdots + \theta_q z^q = 0 \tag{2.1.24}$$

are outside the unit circle. We combine the AR model and the MA model to the $ARMA$ model to make a more sophisticated model (Medium, 2018). The $ARMA$ model can capture both the AR and the MA effects of a time series.

Definition 2.1.9. *Autoregressive Integrated Moving Average Model*

If the data is generated from a non-stationary process, one option is to use an $ARIMA$ time series model to model the data. The $ARIMA$ model can include both non-seasonal and seasonal trends.

A non-seasonal $ARIMA(p, d, q)$ model of order p, d and q for a time series $\{X_t\}_{t=1}^{T}$ is defined as

$$(1 - \sum_{k=1}^{p} \phi_k L^k)(1 - L)^d X_t = (1 + \sum_{k=1}^{q} \theta_k L^k)\epsilon_t$$
$$(1 - L)^d X_t = \sum_{k=1}^{p} \phi_k L^k (1 - L)^d X_t + (1 + \sum_{k=1}^{q} \theta_k L^k)\epsilon_t \tag{2.1.25}$$

where d is a non-negative integer. We have that p is the number of autoregressive terms, d is the number of non-seasonal differences needed for stationarity and q is the number of moving average terms.

An non-seasonal $ARIMA$ model with integration order d reduces to an $ARMA$ model when it is differenced d times. If we set $Y_t = (1 - L)^d X_t$, we have that the time series $\{Y_t\}_{t=1}^{T}$ can be represented by an $ARMA(p, q)$ model

$$Y_t = \sum_{k=1}^{p} \phi_k L^k Y_t + (1 + \sum_{k=1}^{q} \theta_k L^k)\epsilon_t \tag{2.1.26}$$

We can extend the $ARIMA$ model to also include seasonal changes. Seasonal changes in a time series are changes that happen in a regular pattern over S time periods, S is the number of periods until the pattern repeats itself. For quarterly data there are 4 periods in

a season and for monthly data there are 12 periods in a season. For a example, for quarterly data the seasonal difference of Y at time t is $Y_t - Y_{t-4}$.

We denote an $ARIMA$ model with seasonal changes as $ARIMA(p,d,q)\times(P,D,Q)_S$ where $p,\ d,\ q$ and S are defined as above and P is the number of seasonal autoregressive terms, D is the number of seasonal differences and Q is the number of seasonal moving average terms.

Example 2.1.2. *$ARIMA(1,1,1)\times(1,1,1)_4$ model*

Following is an example of the $ARIMA(1,1,1)\times(1,1,1)_4$ model with a seasonal change for a quarterly time series $\{Y_t\}_{t=1}^T$

$$(1-\phi_1 L)(1-\Phi_1 L^4)Y_t = (1+\theta_1 L)(1+\Theta_1 L^4)\epsilon_t. \tag{2.1.27}$$

Definition 2.1.10. *Generalized Autoregressive Conditional Heteroskedasticity Model*

To model the volatility, the second order moment, of a time series the $GARCH$ model is widely used. Engle (1982) first introduced the $ARCH$ model and in his article used the models to estimate the means and variances of the inflation in the U.K.. The $ARCH$ model models the conditional variance as a function of the past squared returns.

Bollerslev (1986) extended the $ARCH$ model to the $GARCH$ model by also including the past conditional variance into the conditional variance term. The $GARCH(p,q)$ model of order p and q for a time series $\{Y_t\}_{t=1}^T$is defined as

$$Y_t = \sigma_t \epsilon_t \tag{2.1.28}$$

$$\sigma_t^2 = \omega + \sum_{i=1}^{p} \alpha_i Y_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2 \tag{2.1.29}$$

where $\omega > 0, \alpha_i, \beta_j \geq 0$ and the innovation sequence, $\{\epsilon_t\}$, is independent and identically distributed with $E(\epsilon_t) = 0$ and $E(\epsilon_t^2) = 1$. $GARCH$ can be used to model and forecast volatility of financial assets among others, and the idea behind the model is that volatility is

persistent and heteroskedastic, i.e. that large values of volatility are likely followed by large values and small are followed by small.

Example 2.1.3. *ARMA - GARCH model*

To model both the mean and the volatility of a time series we can use a combination of the $ARMA$ and $GARCH$ models. The $ARMA$ part capture the mean trend and the $GARCH$ part estimates the volatility. For example $ARMA(1,1) - GARCH(1,1)$ model for a time series $\{Y_t\}_{t=1}^{T}$ is defined as (Song and Kang, 2018)

$$Y_t = \mu + \phi Y_{t-1} + \epsilon_t + \theta \epsilon_{t-1} \tag{2.1.30}$$

$$\epsilon_t = \sigma_t Z_t \ where \ Z_t \sim i.i.d.N(0,1) \tag{2.1.31}$$

$$\sigma_t^2 = \omega + \alpha \epsilon_{t-1}^2 + \beta \sigma_{t-1}^2 \tag{2.1.32}$$

2.2 Forecasting

In a nutshell, forecasting is predicting future outcomes and trends based on historical data, and time series models can be used to predict future values based on past observations (Chatfield, 2000; Investopedia, 2018a). Forecasting is an important tool for predicting future economical outcomes, and evaluating the outcomes is important for making appropriate plans and assisting in the design and implementation of economic policies (Montgomery, Jennings and Kulahci, 2016). Forecasts are produced, studied and evaluated daily by central academia, banks, consumers, firms and practitioners (Rossi, 2014). For example, central banks base their monetary policies on the most likely future paths of key variable such as inflation and exchange rates etc. Firms decide their prices and strategies based on expected, forecasts of sales and financial firms trade based on the forecasts of assets values. Since forecasting is very relevant in different areas in economics and finance, evaluating the forecasts predictive

ability is of much importance. If forecasts are not correctly specified it can lead to poor investments or not the right actions being taken. Following is a brief introduction to three types of forecasts, point, interval and density. We begin with the most simple one, the point forecast, and proceed with an introduction to interval and density forecasts.

2.3 Point Forecast

A point forecast is an estimate for a future value for stochastic variable X, and it is usually its mean or median (Montgomery, Jennings and Kulahci, 2016). The mean forecast is the value μ, defined as $\mu = E(X)$, and the median forecast is the value m, defined as $P(X < m) = 0.5$. The forecast give a guide to immediate action for the forecast user. A bunch of literature has focused on point forecasting and the evaluation methods (Fuller and Hasza, 1981; Engle and Yoo, 1987; Gneiting, 2001). Two of the reasons for this could be that the point forecasts are easy to compute compared to interval- and density forecasts and are easy to understand. Some shortfalls of the point forecast are that the forecast only describes one possible outcome and doesn't give any information about the uncertainty around the forecast such as the interval and density forecasts does.

When the focus is forecasting using time series data, one way to obtain point forecasts as a prediction for the future values is through the use of time-series models such as the $ARMA$ models and the $ARCH$ models. The $ARMA$ models can produce a forecast of the mean or median, and the models depend linearly on the previous data points. The $ARCH$ and $GARCH$ models allow for non-linear dependence in the variance and give a forecast of the data volatility.

To illustrate how to obtain a point forecast through assumptions we will use an $AR(1)$

model: $y_t = \alpha + \beta y_{t-1} + \varepsilon_t$. Let the estimates for α and β obtained at time t be denoted by a_t and b_t, respectively. The parameters are obtained using the full sample of data available up to the time the forecast is made and re-estimated as time goes by and new information is added to the sample. The in-sample fitted errors are obtained from $e_i = y_i - a_t - b_t y_{i-1}$ for $i = 1, 2, \ldots, t$. The forecasted value y_{t+1} based on the information up to time t is obtained as $f_{t+1|t} = a_t + b_t y_t$ and the forecasts are generated as time goes by. Point forecasting can for example be used to forecast interest rates, inflation and stock prices etc.

2.4 Interval Forecast

The interval forecast specifies the probability that a predicted variable will fall within a stated interval. Compared with point forecasts, interval forecasts can be used to assess future uncertainties and plan different strategies for the different possible outcomes (Chatfield, 1993).

An interval forecast consists of upper and lower limits associated with the probability covering the future value, and the limits of an interval forecast are often called prediction bounds (Brockwell and Davis, 1987) or forecast limits (Wei, 1990). The interval is often referred to as a confidence interval (Granger and Newbold, 1986) or a prediction interval (PI) (Abraham and Ledolter, 1983).

There are several ways to calculate an interval forecast and Chatfield (1993) describes several of these methods in his article. One of the methods Chatfield (1993) present is to use the forecast error to obtain an interval forecast, defined as follows: let $\{y\}_{t=1}^{n}$ be an observed time series containing n observations, and we want to point forecast Y_{n+k} conditional on data up to time n for k steps ahead. Let this forecast be denoted by $\widehat{Y}_n(k)$ when it is a random variable and $\widehat{y}_n(k)$ when it is a particular value determined by the observed data.

The conditional forecast error corresponding to $\widehat{y}_n(k)$ is defined as $e_n(k) = \widehat{Y}_n(k) - \widehat{y}_n(k)$. A common form of the PI's is a $(100-\alpha)\%$ for Y_{n+k} is given by

$$\widehat{y}_n(k) \pm z_{\frac{\alpha}{2}} \sqrt{var(e_n(k))} \tag{2.4.1}$$

where $z_{\alpha/2}$ is the appropriate critical value given by the standard normal distribution. Equation (2.4.1) assumes that the forecast is unbiased and that the forecast errors are normally distributed with mean zero and $E(e_n(k)^2) = var(e_n(k))$. Since $\sqrt{var(e_n(k))}$ usually have to be estimated some authors have suggested that $z_{\frac{\alpha}{2}}$ should be replaced by the percentage point of the t distribution with appropriate degrees of freedom (Harvey, 1989).

A common application of interval forecasting is predicting the Value-at-Risk (VaR) for the financial risk for corporations or investment portfolios. VaR is the maximal amount that will be lost with probability p when portfolios are exposed to a random risk (Holton, 2014). VaR is also called the quantile risk measure and is defined as the inverse of a cumulative distribution function for a random risk S, $VaR(S,p) = F_S^{-1}(p)$ (Chiu, Lee and Hung, 2005). For VaR the intervals are one-sided or open intervals. Other applications of interval forecasting in economics and finance are forecasting unemployment rate, gross domestic product growth and stock prices.

2.5 Density Forecast

A density forecast is an estimate of the probability density of predicted variables. The point and interval forecasts can be viewed as by-products of the density forecast, the former being the mean and the latter the quantiles. Forecasts are estimates therefore there are uncertainty around them. Interval forecast represent a first response to point forecasts lack of addressing this uncertainty (Tay and Wallis, 2000). The density forecast is important since it provides a

complete description of the uncertainty around the predicted value and it provides information about how sure the forecaster is regarding the precision around the forecasted value (Bao, Lee and Saltoğlu, 2007). In the recent years, central banks and policy institutions started to realize the importance of measuring and reporting uncertainty around the predicted variables and as a result the demand for density increased (Rossi, 2014). Advances in statistical methodology and increases in computer power also play a part in the increased demand for density forecasts.

The first series of density forecasts that have been produced was made by the Business and Economic Statistics Section of the American Statistical Association (ASA) and the National Bureau of Economic Research ($NBER$) (Tay and Wallis, 2000). The series dates back to 1968 when ASA and $NBER$ combined their efforts and initiated a quarterly survey of macroeconomic forecasters in the U.S. Zarnowitz (1969) explained that the purpose of this survey was to record a suitable record for evaluation for different forecasting assumptions, studies of density forecasts and to analyze the varying degrees of consensus among forecasters. In the U.K. the history of density forecasts dates back to 1992 when the Treasury established the Panel of Independent Forecasters. In addition the panels point forecasts, it was suggested that the panel should report density forecasts for inflation and growth. It took some time before the idea was implemented, and only one set of density forecasts was published from the panel before the panel was dissolved in May 1997. In February 1996 the Bank of England began publishing density forecasts of inflation in its quarterly Inflation Report.

After this we also see a development in finance and more effort was aimed at producing forecasts with a detailed description of the uncertainty for asset and portfolio returns (Raaij and Raunig, 2002). When generating forecasts for financial data the normal distribution is often inadequate, since financial data is known to be conditionally heteroskedastic and unconditionally leptokurtic (Diebold, Gunther and Tay, 1998). When we test for normality, many of the tests rely on the third or the fourth moment and the null hypothesis is rejected if

there is significant skewness and excess kurtosis (Tay and Wallis, 2000). Since we have many examples of empirical studies that discovered non-normal higher moments in the distributions for interest rates, stock returns and other financial data series it suggest that the normal distribution is not the best fit. One example of a study which found evidence of excess kurtosis in a financial data series is Fama (1965). In a study of the daily returns of the stocks listed in the Dow Jones Industrial Average reports evidence of excess kurtosis in the unconditional distributions of the stocks.

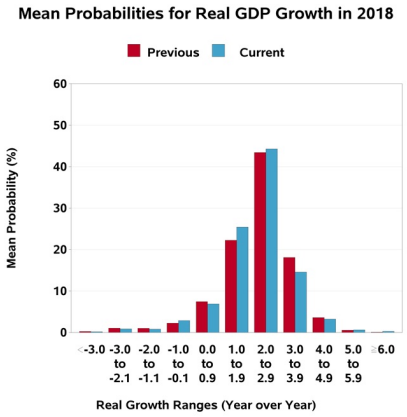
Tay and Wallis (2000) explained that one of many reasons for the interest in more descriptive forecasts in finance is because of risk management. Risk management has grown into a prominent industry and several news agencies and banks regularly issues density forecasts, such as J.P. Morgan, Bloomberg and Reuters among others. The idea behind these forecasts is to have a platform for the user to generate density forecasts of the change in value for customized portfolios over a set holding period. The main target of these density forecasts is to measure the VaR, i.e. the nth percentile of the distribution. If we forecast variables with departures from normality and generate the forecasts inappropriately using the normal assumption, it will affect the usefulness of the VaR estimates. Especially if a portfolio has evidence of excess kurtosis and we have a normality based forecasts, then we will probably underestimate the VaR of the portfolio.

Tay and Wallis (2000) argued that density forecasting in finance began with the literature that aims to model and forecast volatility. Engle (1982) introduced the $ARCH$ model that can be used to model conditional volatility based the squares of the past observations, and this model can produce forecasts with time-varying conditional variances. Compared with the normal distribution, the $ARCH$ model imply larger kurtosis in the unconditional distribution. Since the innovation term in an $ARCH$ model often is assumed to be normally distributed, the $ARCH$ model delivers symmetric density forecasts which is often unsuitable since there

is often evidence of asymmetries in asset returns. The excess kurtosis generated by the $ARCH$ models with normal innovations has been found insufficient to explain the degree of leptokurtosis in many financial time series. Another option for density forecasts for financial time series is the $GARCH$ model. They are often more appropriate since the $GARCH$ model can generate skewed and leptokurtic conditional forecasts. This is accomplished by integrating skew and kurtosis into the distribution of the standardized residuals of the $GARCH$ processes.

Three ways to obtain a density forecast are through a forecast survey, relying on assumptions about the error term and through stochastic simulations. In a survey forecast, the survey respondents give an answer to how likely the target variable will lie between certain percentage points and will directly provide the density distribution and the quantiles. The survey forecasts can be presented as a histogram (Rossi, 2014). An example of a density survey forecast is the Survey of Professional Forecasters (SPF). It is based on data from the Federal Reserve Bank of Philadelphia and is a quarterly survey of macroeconomic forecasts. The survey forecasts real gross domestic product (GDP) growth, unemployment rate and consumer price index. Figure 2.1 below is an example of the density survey forecasts from the SPF. The example is a density forecast of the GDP growth for 2018 and it was produced in the third quarter of 2017. The red bars is the forecast made in the previous quarter.

Figure 2.1: SPF Forecast of U.S. Real GDP growth rate made in Q3 2017



An example of the third quarter 2017 Survey of Professional Forecasters. Reprinted from Federal Resevere Bank of Philadelphia. Retrieved 21. Aug 2017 from $https://www.philadelphiafed.org/research-and-data/real-time-center/survey-of-professional-forecasters/2017/survq317$

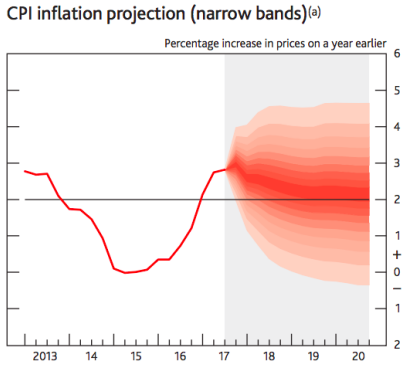
To illustrate how to obtain a density forecast through assumptions on the error term we will use the $AR(1)$ model from the point forecast example. The first step is to estimate a variance based on the in sample fitted errors, $\widehat{e_i} = y_i - a_t - b_t y_{i-1}\ for\ i = 1, 2, \ldots, t$ (Rossi, 2014). Typically, we assume that the error term is normally distributed with mean zero. Then we can produce a density forecast by assuming that the point forecast up to time t, $f_{t+l|t}$, is normally distributed with mean $a_t + b_t y_t$, and use the estimated variance to estimate the density around the mean.

A third option for generating density forecasts are through stochastic simulation methods. Based on a model for a time series we can simulate future paths for the variable. The future paths can then be used to estimate the mean, median and quantiles for the time series.

One example of density forecasting is to use it to forecast inflation. Bank of England has published a density forecast of inflation since 1996 and presented it as a fan chart. The density forecast is represented graphically as a set of PI covering the $10\%, 20\%, \ldots, 90\%$ of the

probability distribution, of lighter shades for the outer bounds. Equivalently, the boundaries for the bands are the $5th, 10th, \ldots, 95th$ percentiles of the density forecasts.

Figure 2.2: Bank of England Fan Chart for Inflation from Nov 2017



An example of the Bank of England fan charts for inflation. Reprinted from the Bank of England. Retrieved 8. Feb 2018 from

$https://www.bankofengland.co.uk/-/media/boe/files/inflation-report/2017/fan-charts-nov-2017.pdf?la=en\&hash=159F9B672EB07084AC010113B657C4B2487F63DB$

Chapter 3

Evaluation Methods Corresponding to Point and Interval Forecasting

Evaluating forecasts is important for determining the accuracy of a forecast. Good quality forecast leads to good decisions, and since forecasting is widely used it is important to evaluate the quality. Evaluating forecast can have various purposes. For example verify if point forecasts are, one average, hitting the predicted outcome. In the context of predicting the Value-at-Risk interval, it could establish if the model used has the right coverage probability.

As in the previous section, we will start with a discussion of evaluation of point and interval forecasts. We will start with a brief introduction to point forecast evaluation. Then we will give a more detailed description of a method for interval forecast evaluation since it is a basis for one method for density forecast evaluation. In the next chapter, we go through a detailed description of a three different density forecast evaluation methods since it is our main focus.

3.1 Point Forecast Evaluation

There are several ways to evaluate point forecasts; one is to test if a model satisfies certain desirable qualities and another is to make comparisons among forecasting models (Rossi, 2014). To illustrate the former evaluation method, we will use a forecasting model, e.g. an $AR(1)$ model, and show how to evaluate if the model satisfy desirable qualities. One of the desirable properties is that the forecast is unbiased. An estimator $\widehat{\theta}$ is unbiased if the expected value of the estimator is equal to the true value of the parameter being estimated, $E(\widehat{\theta}) = \theta$. For a point forecast this is equal to $E(y_t) = E(a_t + b_t y_{t-1} + e_t) = f_{t|t-1}$, since the error term is assumed to have zero mean. The forecast might over- and under predict its target value but on average the model will do a good job, and the over- and under predictions will cancel each other out. One way to test if the forecast is unbiased is by evaluating if the forecast error is zero on average. We can use the one-step ahead forecast error model, $e_{t+1|t} = y_{t+1} - f_{t+1|t}$, to test for unbiasedness of the forecast. We can do this by regressing the forecast error on a constant and test if the forecast error has mean zero. We then estimate the following regression $e_{t+1|t} = \theta + u_{t+1,t}$ where $u_{t+1,t}$ is the error in the regression. Based on a t-test we can determine if the forecast error is zero on average. If we reject the hypothesis that θ is zero, then we conclude that the forecast is biased.

The second desirable property of the forecast errors is that are not supposed to be predictable based on the data available at the time (Rossi, 2014). I.e. if we included all the available data we assume to be significant we should not be able to predict the forecast errors. If the errors are predictable, we need to include more data in the model which will improve the forecast. We can evaluate if errors are predictable using a test for forecast rationality. First we investigate our data and find a variable we think could be a useful predictor. Let z_{i-1} denote the omitted variable in our AR model. The next step is to estimate the following

regression: $e_{t+1|t} = \theta_1 + \theta_2 z_t + u_{t+1,t}$, where $u_{t+1,t}$ denotes the error in the regression. We can determine if the forecast errors are zero on average with a joint significant test with null hypothesis: $H_0 : \theta_1 = \theta_2 = 0$. If the forecast errors are zero an average, then θ_1 and θ_2 should be jointly zero. If we accept the null hypothesis, we conclude that the forecast is rational.

Rossi (2014) illustrated an example of another method to evaluate point forecasts. The method is to check the models relative forecasting ability, and the objective is to evaluate if the forecasting ability for two or more models is similar. First step is to choose two or more different forecasting models then investigate is one has a significantly larger expected loss function. Examples of loss functions are the mean squared forecast error ($MSFE$) and mean absolute error (MAE). The $MSFE$ of a model is defined as: $MSFE = P^{-1} \sum_{t=r}^{T} e_{t+1|t}^2$ where P is the number of forecasts. If we want to compare models based on the $MSFE$ we compare the models based on their relative average squared values of the forecast error. We can do this by calculating the difference in the $MFSE$'s for the two models and evaluate if the difference is zero using a t-test. If the difference is zero the forecasting abilities are similar.

3.2 Interval Forecast Evaluation

The traditional method to evaluate interval forecast only take into account the coverage and the first moment, and test if the forecast is correct on average (Intensity, 2017). When using the traditional method, we first make a track record of interval forecasts and the outcomes related to the forecasts of a sample of points in time. We can use our observed track record of the forecasts and related outcome to evaluate the forecasting model. The larger the sample we have observed, the better our evaluation of our model will be. Say we have a track record for a $(100 - \alpha)\%$ interval forecast and the outcome of the corresponding forecasted variable, then we expect the outcome to fall inside the interval forecast $(100 - \alpha)\%$ of the time and outside

α% of the time. Suppose we observe a time series and the corresponding interval forecasts with $\alpha = 0.1$. If more than the 10% of the outcomes fall outside the interval forecasts, the forecast is too narrow on average. Conversely, if the outcomes fall outside less than 10% of the time, the forecast in too wide on average. A formal test for this is Christoffersen test for unconditional coverage, explained in detail later in this section.

One large shortfall of the traditional method is that it ignores the information from higher-order dynamics, such as the volatility. The information from the second moment is crucial when it comes to forming dynamic interval forecasts, recognizing that the interval should be narrow in tranquil times and wide in volatile times (Engle, 1982). A result of this is that an interval forecast that doesn't take into account higher-order dynamics is that it may be correct on average, but in periods it will have incorrect conditional coverage, characterized by clusters of outliers. In order to build a test to evaluate the conditional coverage, Christoffersen (1998) introduced a framework that combines 3 tests to evaluate interval forecasts: a likelihood ratio (LR) test for unconditional coverage, a LR test for independence and a joint LR test for coverage and independence. The tests are based on a testing criterion that is defined as follows. Let $\{y\}_{t=1}^{T}$ denote an observed sample path of the time series y_t and the out-of-sample interval forecast, $\{L_{t|t-1}(p), U_{t|t-1}(p)\}_{t=1}^{T}$, corresponding to the series is also available. $L_{t|t-1}(p)$ and $U_{t|t-1}(p)$ are the upper and lower limits of ex ante interval forecast for time t at time $t-1$ for the coverage probability p.

We can define an indicator variable based on the realizations of the time series and the interval forecasts. The indicator variable, I_t, for a given interval forecast $(L_{t|t-1}(p), U_{t|t-1}(p))$ at time t made at time $t-1$, is defined as

$$I_t = \begin{cases} 1 \; if \; y_t \in [L_{t|t-1}(p), U_{t|t-1}(p)] \\ \\ 0 \; if \; y_t \in [L_{t|t-1}(p), U_{t|t-1}(p)] \end{cases} . \tag{3.2.1}$$

With the indicator variable Christoffersen (1998) established a general testing criterion for interval forecast defined as: We define the sequence of interval forecasts $\{L_{t|t-1}(p), U_{t|t-1}(p)\}_{t=1}^{T}$ as efficient with respect to information set Ψ_{t-1} if $E[I_t|\Psi_{t-1}] = p \; \forall \; t$. Then we can test if the interval forecast is efficient by testing if $E[I_t|\Psi_{t-1}] = p \; \forall \; t$.

Letting the information set consists of the past realizations of the indicator sequence, $\Psi_{t-1} = \{I_{t-1}, I_{t-2}, \ldots, I_1\}$, then $E[I_t|\Psi_{t-1}] = E[I_t|I_{t-1}, I_{t-2}, \ldots, I_1] = p \; \forall \; t$. Then testing $E[I_t|\Psi_{t-1}] = p \; \forall \; t$ is equivalent to testing if $\{I_t\} \sim$ *i.i.d. Bernoulli*(p), and the sequence of interval forecasts $\{L_{t|t-1}(p), U_{t|t-1}(p)\}_{t=1}^{T}$ has correct conditional coverage if $I_t \sim$ *i.i.d. Bernoulli*$(p) \; \forall \; t$.

To test for unconditional coverage Christoffersen (1998) suggested using a LR test and the test is based on the indicator sequence $\{I_t\}_{t=1}^{T}$ constructed from a given interval forecast. To test for unconditional coverage, the null hypothesis $E[I_t] = p$ should be tested against the alternative hypothesis that $E[I_t] \neq p$ given that the sequence is independent. Under the null hypothesis the likelihood function is $L(p; I_1, I_2, \cdots, I_T) = (1-p)^{n_0} p^{n_1}$ and under the alternative is $L(\pi; I_1, I_2, \cdots, I_T) = (1-\pi^{n_0})\pi^{n_1}$. Then testing for unconditional coverage can be done with a standard likelihood ratio test;

$$LR_{uc} = -2log\frac{L(p; I_1, I_2, \ldots, I_T)}{L(\widehat{\pi}; I_1, I_2, \ldots, I_T)}$$

$$= 2[n_0 log(\frac{1-p}{1-\pi}) + n_1 log(\frac{p}{\pi})] \sim asymptotically \; \chi^2(s-1) = \chi^2(1) \qquad (3.2.2)$$

where the maximum likelihood estimate (MLE) for π is $\widehat{\pi} = n_1/(n_1 + n_0)$ and the number of possible outcomes, s, in the sequence is 2. Out of n observation, we let n_1 denote the number of outcomes falling inside the corresponding interval forecast, let n_0 denote the number of outcomes that fall outside and $n = n_1 + n_0$.

The next test for interval forecast is to check for independence. The independence hypothesis is tested against a first-order Markov chain alternative. First consider a binary first-order

Markov chain, $\{I_t\}$, with transition probability matrix

$$\Pi_1 = \begin{bmatrix} 1-\pi_{01} & \pi_{01} \\ 1-\pi_{11} & \pi_{11} \end{bmatrix} \tag{3.2.3}$$

where $\pi_{ij} = P(I_t = j | I_{t-1} = i)$. Then the approximate likelihood function for the process is

$$L(\Pi_1; I_1, I_2, \ldots, I_T) = (1-\pi_{01})^{n_{00}} \pi_{01}^{n_{01}} (1-\pi_{11})^{n_{10}} \pi_{11}^{n_{11}} \tag{3.2.4}$$

where n_{ij} is the number of observations in the state i followed by the state j. Conditioned on the first observations everywhere the parameters for the maximized log-likelihood functions are simply the rations of counts for the appropriate cells:

$$\widehat{\Pi}_1 = \begin{bmatrix} \frac{n_{00}}{n_{00}+n_{01}} & \frac{n_{01}}{n_{00}+n_{01}} \\ \frac{n_{10}}{n_{10}+n_{11}} & \frac{n_{11}}{n_{10}+n_{11}} \end{bmatrix}. \tag{3.2.5}$$

Then consider the output sequence, $\{I_t\}$, from an interval model and estimate a first-order Markov chain model on the sequence. To test the hypothesis that the sequence is independent noting that the transition probability matrix

$$\Pi_2 = \begin{bmatrix} 1-\pi_2 & \pi_2 \\ 1-\pi_2 & \pi_2 \end{bmatrix} \tag{3.2.6}$$

corresponds to independence. Then the likelihood function under the null hypothesis is $L(\Pi_2; I_1, I_2, \ldots, I_T) = (1-\pi_2)^{(n_{00}+n_{10})} \pi_2^{(n_{01}+n_{11})}$ and the MLE for Π_2 is $\widehat{\Pi}_2 = \widehat{\pi}_2 = (n_{01} + n_{11})/(n_{00} + n_{10} + n_{01} + n_{11})$. Then the LR test is

$$LR_{ind} = -2log \frac{L(\widehat{\Pi}_2; I_1, I_2, \ldots, I_T)}{L(\widehat{\Pi}_1; I_1, I_2, \ldots, I_T)} \sim asymptotically\ \chi^2((s-1)^2) = \chi^2(1) \tag{3.2.7}$$

since s denotes the number of possible outcomes and the sequence is binary, we have $s = 2$. The test doesn't depend on the true coverage, p, therefore it only tests for independence.

The last test that Christoffersen (1998) introduced was a test for conditional coverage as a combination of the unconditional coverage and independence tests. If, conditioned on the

first observation in the test for unconditional coverage, then $\widehat{\pi} = \widehat{\pi}_2 = \widehat{\Pi}_2$. This implies, if the first observation is ignored, then the test for conditional coverage is $LR_{cc} = LR_{uc} + LR_{ind}$, where

$$LR_{cc} = -2log\frac{L(p; I_1, I_2, \ldots, I_T)}{L(\widehat{\Pi}_1; I_1, I_2, \ldots, I_T)} \sim asymptotically\ \chi^2(s(s-1)) = \chi^2(2) \tag{3.2.8}$$

since $s = 2$. This test enables joint testing of randomness and correct coverage, while retaining the individual hypotheses as subcomponents. The three tests Christoffersen (1998) presented can be applied in a natural sequence when evaluating interval forecasts. The first step is use the joint test for goodness of fit and independence LR_{cc}. If we accept the null hypothesis we can conclude that we have specified the correct interval forecasts model and we don't have clusters of outliers. If the null hypothesis is rejected we can investigate why with the LR_{uc} and LR_{ind} tests. Is it because the model failed to capture time dependence or because the model doesn't have the correct coverage?

The three tests Christoffersen presented can also be evaluated using the Pearson chi-squared statistic (Wallis, 2003). One standard way of denoting the Pearson Chi squared statistic is

$$\sum \frac{(O-E)^2}{E} \tag{3.2.9}$$

where O denotes the observed probabilities and the E denotes the expected probabilities (DeGroot and Schervish, 2012). The equivalent chi-squared statistics for the unconditional coverage test is

$$X^2 = \frac{n(p-\pi)^2}{\pi(1-\pi)}. \tag{3.2.10}$$

For the independence test we first denote the matrix for the for the observed frequencies as

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \tag{3.2.11}$$

and then the corresponding chi-squared statistics is

$$X^2 = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}. \tag{3.2.12}$$

For the joint test for coverage and independence we denote the observed and expected frequencies as, respectively:

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \; and \; \begin{bmatrix} (1-\pi)(a+b) & \pi(a+b) \\ (1-\pi)(c+d) & \pi(c+d) \end{bmatrix}. \tag{3.2.13}$$

Each of the Pearson's chi-squared statistics can be evaluated by comparing them to the appropriate critical values from the χ^2 distribution with the same degrees of freedom as the corresponding LR tests for interval forecasts.

Wallis (2003) mentioned in his paper that Christoffersen's method for evaluating interval forecasts could be extended for density forecasts and he had an introduction to this extension. He stated that for density forecasts the first question raised is for goodness of fit of the forecasts. Two classical methods in statistics for measuring goodness of fit is the likelihood ratio and the Pearson chi squared test. The first step to extend the method to test for goodness of fit to be valid for density forecasts is to divide the interval range of the predicted variable into k mutually exclusive classes and analyze if the probabilities of outcomes in each of the classes for the forecast densities are similar to the observed relative frequencies. In formal terms, we divide the variable range into k mutually exclusive classes and let n_i denote the outcomes that fall into each of these classes, $i = 1 \ldots k$ and $\sum n_i = n$. We have the z-transform for a density forecast defined as $z = F(y)$ where y denotes the observed outcome, $F(.)$ denotes the distribution function for the density and z denotes the forecast probability of observing an outcome no greater than the actual realized value. The argument used for the density forecasting method presented by Diebold, Gunther and Tay (1998) is if $F(.)$ is correct then z has a $U(0,1)$ distribution. We can use the z-transformation for our variable range in

the following manner, the z-transforms can be divided into classes with boundaries j/k where $j = 0, 1, \ldots k$. Then we can implement the Pearson chi-squared statistics for goodness-of-fit as

$$X^2 = \sum \frac{(n_i - \frac{n}{k})^2}{\frac{n}{k}} \tag{3.2.14}$$

and the likelihood ratio test statistics is

$$LR = 2 \sum n_i log(\frac{kn_i}{n}). \tag{3.2.15}$$

Under the null hypothesis, both of the X^2 and the LR is distributed as χ^2 with $k-1$ degrees of freedom. Li and Andersson (2018) further studied the idea of extending the evaluation tests for interval forecasting to be valid for evaluating density forecasts and the method will be explained further in the next chapter.

Chapter 4

Evaluation Methods Corresponding to Density Forecasting

Until recent years, little attention has been given to the evaluation methods for density forecasts. The are several different factors that can be the cause of this neglect (Diebold, Gunther and Tay, 1998). One of the factors is that historically the construction of density forecasts has required quite restrictive, even dubious, assumptions such as Gaussian innovations, linear dynamics and no parameter estimation uncertainty. Because of recent work using numerical and simulation techniques and improvements in data technology we don't have to rely on these assumptions to the same degree and have made it easier to provide credible density forecasts.

Another factor to why density forecasts haven't been produced and evaluated is because of the demand. In the past, point and interval have usually given enough information for the most users need. However, due to recent developments the situation has changed, especially in quantitative finance, and the demand has increased. In areas like financial risk management, they are completely dedicated to providing density forecasts of portfolio values and to tracking certain parts of the densities, for example Value-at-Risk.

A third reason is that the evaluation of density forecasts appear to be a difficult task. It is possible to modify techniques for evaluating point and interval forecasts and use them for evaluating density forecasts, but this application can lead to incomplete evaluation of density forecasts. We could use Christoffersen's (1998) method for evaluating interval forecasts without extending it. We could use this method to check if the series of 95% PI corresponding to a series of density forecasts are accurately conditionally calibrated. The issue with this type of evaluation is that it leaves the question if the corresponding PI at other confidence levels are accurately conditionally calibrated unanswered. To determine if we have the correct conditional calibration of density forecasts, all the PI must be simultaneously be correctly conditionally calibrated for all possible confidence levels.

Because of the increase of demand of density forecasts, the demand for proper evaluation methods has also increased and we will give a detailed description of different methods of evaluating density forecasts. Density forecasts can be evaluated using the probability integral transform (PIT), the loss function and the standard uniform distribution. A PIT is the cumulative probability evaluated at the actual, realized value of a target variable. The PIT measures the likelihood of observing a value less than the actual realized value, where the probability measured is the density forecast. Diebold, Gunther and Tay (1998) showed that if a PIT is $i.i.d.$ $U(0,1)$ if the density forecast is correct. Then to evaluate if the density forecast is correctly specified is the same as testing if the PIT is $i.i.d.$ $U(0,1)$.

There is also another method based on the PIT, combining it with the normal distribution and likelihood ratio test. The test is based on transforming the PIT using the inverse distribution function of a standard normal, then calculating the log-likelihood and constructing the likelihood ratio test statistic.

Another method is to extend the Christoffersen testing framework using the Pearson's

chi-squared based statistics, multinomial distribution and Markov chains. The extension is a non-parametric test and no parametric model is needed for the independence test. The test can evaluate the goodness of fit and independence at the same time. These three evaluation methods will be explained in more details later in this chapter.

4.1 Loss Functions and Action Choices

Since the problem of evaluating density forecasts can be linked to the forecast user's loss function, we will give a short illustration between the relationship between density functions, loss functions and action choices (Diebold, Gunther and Tay, 1998). First, we define the decision environment as follows: Let $\{f_t(y_t|\Omega_t)\}_{t=1}^n$ be a sequence of conditional densities dependent on a series y_t, where $\Omega_t = \{y_{t-1}, y_{t-2}, \dots\}$. Let $\{p_t(y_t|\Omega_t)\}_{t=1}^n$ be a corresponding sequence of 1-step-ahead density forecasts and let $\{y_t\}_{t=1}^n$ denote the corresponding series of realizations. We denote the forecaster user loss function as $L(a, y)$ where a denotes an action choice, and the forecast user chooses an action to minimize the expected loss computed using the density forecast, $p(y)$. Assuming that $p(y)$ being the correct density, the forecasts user chooses an action $a^*(p(y)) = \underset{a \in A}{argmin} \int L(a, y)p(y)dy$ where A denotes all the choices the forecast user might make. The choice a^* incurs a loss $L(a^*, y)$ and this loss is a random variable. The expected loss with respect to the true density $f(y)$ is defined as $E[L(a^*, y)] = \int L(a^*, y)f(y)dy$. Different density forecasts will, in general, generate different action choices, therefore also different distributions of loss. The closer the density forecast is to the true data generating process, the lower is its expected loss.

Different forecasts can lead to different actions that minimizes the loss and different ranking of forecasts. Suppose a forecast user has the option between choosing two different forecasts,

denoted by $p_j(y)$ and $p_k(y)$ where j and k denotes the different forecasts. Then the user will prefer forecast $p_j(y)$ to $p_k(y)$ if $E[L(a_j^*, y)] \leq E[L(a_k^*, y)]$, where a_j^* denotes the action that minimized the expected loss based on the forecast j. Ideally, we want a ranking of forecasts which all the forecasts users agree with, not dependent on their loss functions. Such a ranking does not exist for incorrect density forecast, i.e. there does not exist a ranking r for arbitrary density forecasts p_j and p_k, both different from the true data generating process f, such that for all loss functions $L(a, y)$ we have

$$r_j \geq r_k \Leftrightarrow \int L(a_j^*, y) f(y) dy \geq \int L(a_k^*, y) f(y) dy. \tag{4.1.1}$$

This can be illustrated by finding to loss functions L_1 and L_2, a density function f governing y, and two density forecasts, p_j and p_k, such that $E[L_1(a_j^*, y)] < E[L_1(a_k^*, y)]$ while $E[L_2(a_j^*, y)] > E[L_2(a_k^*, y)]$. In other words, the forecast user with loss function L_1 does better on average under forecast j, and the forecast user with loss function L_2 does better on average under forecast k.

The illustration show that is no way to rank to incorrect density forecasts such that all the users will agree with the ranking, but suppose we have a forecast that agree with the true data generating process. All forecasts users, regardless of loss function, will prefer this forecast. Statistically, suppose that we have $p_j(y) = f(y)$, so that the action a_j^* minimizes the expected loss corresponding to the true data generating process. Then $\int L(a_j^*, y) f(y) dy \leq \int L(a_k^*, y) f(y) dy \; \forall \; k$, since a_j^* minimizes the expected loss over all possible actions. The insight that $f(y)$ dominates all other forecasts all for all users, regardless of loss function, suggests a useful direction for evaluating density forecasts. The evaluation of density forecasts boils down to evaluating if $\{p_t(y_t|\Omega_t)\}_{t=1}^n = \{f_t(y_t|\Omega_t)\}_{t=1}^n$.

4.2 Evaluating Density Forecasts with PIT and Uniform Distribution

Determining if $\{p_t(y_t|\Omega_t)\}_{t=1}^n = \{f_t(y_t|\Omega_t)\}_{t=1}^n$ appear to be a difficult task since we can observe y_t but never observe the true distribution $\{f_t(y_t|\Omega_t)\}_{t=1}^n$ at time t. Another observation that also makes the evaluation difficult is that $f_t(y_t|\Omega_t)$ may display structural change, as indicated by its time subscript. Even though the task is challenging, we still have certain methods for evaluating density forecasts.

Diebold, Gunther and Tay (1998) suggested a way to evaluate density forecasts that is based on the relationship between the data generating process, $f_t(y_t)$, and the sequence of density forecasts, $p_t(y_t)$. The relationship between the data generating process and the density forecast is related through the probability integral transform, z_t, of the realization of the process taken with respect to the density forecast. The PIT is the distribution function corresponding to the density $p_t(y_t)$ evaluated at $y_t, z_t = \int_{-\infty}^{y_t} p_t(u)du = P_t(y_t)$. Let denote the density for z_t as $q_t(z_t)$. This density is significant when it comes to the evaluation. If we assume that $\frac{\partial P_t^{-1}(z_t)}{\partial z_t}$ is non-zero and continuous over the support y_t and by recognizing that $P_t^{-1}(z_t) = y_t$ and $p_t(y_t) = \partial P_t(y_t)/(\partial y_t)$, then we can see that z_t has support on the unit interval with density

$$q_t(z_t) = |\frac{\partial P_t^{-1}(z_t))}{\partial z_t}| f_t(P_t^{-1}(z_t)) = \frac{f_t(P_t^{-1}(z_t))}{p_t(P_t^{-1}(z_t))} \tag{4.2.1}$$

since

$$\begin{aligned} p_t(y_t) = \frac{\partial P_t(y_t)}{\partial y_t} &\Rightarrow p_t(P_t^{-1}(z_t)) = \frac{\partial P_t(P_t^{-1}(z_t))}{\partial y_t} \\ \Rightarrow \frac{\partial y_t}{\partial z_t} = \frac{1}{p_t(P_t^{-1}(z_t))} &\Rightarrow \frac{\partial (P_t^{-1}(z_t))}{\partial z_t} = \frac{1}{p_t(P_t^{-1}(z_t))}. \end{aligned} \tag{4.2.2}$$

Note that $q_t(z_t)$ is distributed as $U(0,1)$ if $p_t(y_t) = f_t(y_t)$. The next step is to extend the

one-period characterization of the density of z_t to characterize the density and dependence structure of the entire z_t sequence, in the case of $p_t(y_t) = f_t(y_t)$.

Suppose we have a sequence, $\{y_t\}_{t=1}^n$, generated from $\{f_t(y_t|\Omega_t\}_{t=1}^n$ where $\Omega_t = \{y_{t-1}, y_{t-2}, \dots\}$. Then, if we have a sequence of density forecasts $\{p_t(y_t|\Omega_t)\}_{t=1}^n$ that coincides with $\{f_t(y_t|\Omega_t\}_{t=1}^n$ and under the usual condition of a nonzero Jacobian with continuous partial derivatives, the sequence of PITs of $\{y_t\}_{t=1}^n$ with respect to $\{p_t(y_t|\Omega_t)\}_{t=1}^n$ is *i.i.d.* $U(0,1)$. In other words, $\{z_t\}_{t=1}^n \sim$ *i.i.d.* $U(0,1)$.

This property can be illustrated with the following steps. The joint density for the sequence $\{y_t\}_{t=1}^n$ can be expressed as $f(y_n, \dots, y_1|\Omega) = f_n(y_n|\Omega_n)f_{n-1}(y_{n-1}|\Omega_{n-1}) \dots f_1(y_1|\Omega_1)$. Then the joint density of $\{z_t\}_{t=1}^n$ can be computed using the change of variables formula:

$$\begin{aligned} q(z_n, \dots, z_1) &= \begin{vmatrix} \frac{\partial y_1}{\partial z_1} & \dots & \frac{\partial y_1}{\partial z_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_n}{\partial z_1} & \dots & \frac{\partial y_n}{\partial z_n} \end{vmatrix} f_n(P_n^{-1}(z_n)|\Omega_n)f_{n-1}(P_{n-1}^{-1}(z_{n-1})|\Omega_{n-1}) \dots f_1(P_1^{-1}(z_1)|\Omega_1) \\ &= \frac{\partial y_1}{\partial z_1}\frac{\partial y_2}{\partial z_2} \dots \frac{\partial y_n}{\partial z_n} f_n(P_n^{-1}(z_n)|\Omega_n)f_{n-1}(P_{n-1}^{-1}(z_{n-1})|\Omega_{n-1}) \dots f_1(P_1^{-1}(z_1)|\Omega_1) \end{aligned} \tag{4.2.3}$$

since the Jacobian transformation matrix is lower triangular. Then we have

$$q(z_n, \dots, z_1|\Omega) = \frac{f_n(P_n^{-1}(z_n)|\Omega_n)}{p_n(P_n^{-1}(z_n))} \frac{f_{n-1}(P_{n-1}^{-1}(z_{n-1})|\Omega_{n-1})}{p_{n-1}(P_{n-1}^{-1}(z_{n-1}))} \dots \frac{f_1(P_1^{-1}(z_1)|\Omega_1)}{p_1(P_1^{-1}(z_1))}. \tag{4.2.4}$$

Then, under the assumptions, each of the ratios above is distributed as $U(0,1)$ and the product of the ratios gives an $n-variate$ $U(0,1)$ distribution for $\{z_t\}_{t=1}^n$. Since the joint distribution is the product of the marginals, we have that $\{z_t\}_{t=1}^n \sim$ *i.i.d.* $U(0,1)$.

The theory suggests that we can evaluate density forecasts by assessing whether the PIT series $\{z_t\}_{t=1}^n$ is *i.i.d.* $U(0,1)$ or not. One example of test we can use is the Kolmogorov-Smirnov (KS) test (Chakravarti, Laha and Rot, 1967). One shortfall of this evaluation

method is that when the rejection occurs there is no indication to why. In other words, say the KS test rejects the hypothesis of $\{z_t\}_{t=1}^n$ is $i.i.d.$ $U(0,1)$, is it because of violation of unconditional uniformity, violation of $i.i.d.$. or both? Even if we could tell why the hypothesis was rejected, we would like to know more. I.e. if we know the rejection comes from violation of uniformity, we would want to know precisely the nature of the violation of uniformity and how important it is, or if we know the rejection comes from violation of $i.i.d.$, what precisely is its nature? Is z dependent or is z heterogeneous but independent?

To make up for the shortfalls Diebold, Gunther and Tay (1998) suggested to use less formal, graphical tools to supplement the formal tests. First, with regard to the unconditional uniformity we can a graphical tool such as the histogram. Using histograms will do the trick since they allow for straightforward imposition of the constraint that z has support on the unit interval. Then we can visually compare the estimated density to the $U(0,1)$ distribution, and we can compute the confidence interval under the null hypothesis of $i.i.d.$ $U(0,1)$, taking advantage of the binomial structure, bin by bin.

Second, with regard to the $i.i.d.$ part of the hypothesis, we can use another graphical tool, the correlogram supplemented with confidence intervals. We can use the correlogram to detect dependence patterns in z. If we detect patterns of correlation in z, it implies that they are not independent. The converse is not always true. We can not conclude that a data set is independent purely based on uncorrelatedness even though we can conclude that a data set is not independent based on lack of uncorrelatedness.

4.3 Evaluating Density Forecasts with Likelihood Ratio and Standard Normal Distribution

Berkowitz (2001) developed a method for evaluating density forecasts based on Rosenblatt (1952) transformation of realizations of variables into a series of *i.i.d.* random variables and likelihood ratio test. The Rosenblatt transformation is defined as

$$x_t = \int_{-\infty}^{y_t} \widehat{f}(u)du = \widehat{F}(y_t) \tag{4.3.1}$$

where y_t is the ex post realization of a variable and $\widehat{f}(.)$ is the ex ante forecasted density. With this transformation, x_t is *i.i.d.* $U(0,1)$, and forecasts users can operate with the forecasts distribution $\widehat{F}(.)$ this and test for violations in independence and of uniformity. A shortfall of this evaluation method is that a large sample size is needed to test for uniformity, and Berkowitz (2001) demonstrated that for sample sizes under a 1000 the test based on uniform transformation showed low power.

Since it is difficult to test for uniformity with small data samples, Berkowitz (2001) introduced an extension of the Rosenblatt transformation that transform the realizations into *i.i.d.* $N(0,1)$ variates under the null hypothesis. This transformation makes it possible for estimation of the Gaussian likelihood, then construct the LR, Lagrange Multiplier (LM) or Wald statistic. Berkowitz (2001) chose to focus on the LR test since it is the uniformly most powerful (UMP) test for some classes of model failure. A UMP test has higher power compared to other test for a fixed confidence level for every value of the unknown parameter (Casella and Berger, 1990). In some cases we cannot show that a test is UMP, but even when that is the case the LR test often has desirable statistical properties and good finite-sample behavior (Hogg and Craig, 1965). Another reason to why focus on the LR test instead of LM or Wald statistics is that the researcher has broad range of how many and which restrictions

to test.

The extension of the Rosenblatt transformation is defined as follows: Let ϕ^{-1} denote the inverse of a standard normal distribution, $N(0,1)$. The following result holds for any sequence and does not depend on the underlying distribution of portfolio returns.

If we a have a series $x_t = \int_{-\infty}^{y_t} f(u)du$ that is *i.i.d.* $U(0,1)$ distributed, then

$$z_t = \Phi^{-1}[\int_{-\infty}^{y_t} f(u)du] \; is \; an \; i.i.d. \; N(0,1). \tag{4.3.2}$$

The transformation is well known and is used to for simulating random variates. From the PIT we know that if we have X, a continuous random variable, with cumulative distribution function F_X, then the r.v. $Y = F_X(X)$ is distributed as a standard uniform random variable. The inverse probability integral transform states that if we have Y, X and F_X, then $F_X^{-1}(Y) = X$ has the same distribution as X. In our case $Y = \int_{-\infty}^{y_t} f(u)du$, $X = z_t$ and $F_X^{-1} = \Phi^{-1}$, which implies we can transform the observed portfolio relazations to a series that should be independent and identically distributed standard normal variables, $z_t = \Phi^{-1}[\widehat{F}(y_t)]$. This is very useful due to the fact that under the null hypothesis the data follow a normal distribution, and this allows us to use the convenient tools associated with the Normal likelihood.

In his paper, Berkowtiz (2001) showed that the following property holds: let $h(z_t)$ denote the density of z_t and $\phi(z_t)$ denote the density of a standard normal, then we have the property

$$log[\frac{f(y_t)}{\widehat{f}(y_t)}] = log[\frac{h(z_t)}{\phi(z_t)}]. \tag{4.3.3}$$

The property establishes inaccuracies in the density forecasts will be retained in the transformed data, i.e. if $f > \widehat{f}$ then the same holds for $h(z_t) > \phi(z_t)$.

None of the transformations, Rosenblatt or to normality, is restricted to distributional assumptions of the underlying data. Thus, if we have the correct density forecast implies that

we have normality of the transformed variables. This means we can test for non-normality and serial correlation for z_t.

One shortfall of the LR test is that it only has power to detect non-normality through the first and second moment of the distribution. If we have correctly specified the conditional first and second moment then the likelihood function is maximized at their true values (Bollerslev and Wooldridge, 1992). If the LR test fails to reject the null, a solution is to test the standardized z_t to non-parametric test. Another solution is to use the LM test framework instead of the LR test framework, then it might be possible to test for normality simultaneously with other restrictions under the null hypothesis.

The basic testing framework is as follows, suppose we have generated the sequence $z_t = \Phi^{-1}(\widehat{F}(y_t))$ for a given model. The sequence z_t should be normal so we have many possible tests that can be constructed. One example that we will explore more in the next chapter, is that the null hypothesis can be tested against a first-order AR alternative with possible values for the mean and the variance different from 0 and 1.

We can write the model as $z_t - \mu = \rho(z_{t-1} - \mu) + \epsilon_t$. The null hypothesis is then $\mu = 0$, $\rho = 0$ and $var(\epsilon_t) = 1$ The log-likelihood equation for the model is

$$-\frac{1}{2}log(2\pi) - \frac{1}{2}log\frac{\sigma^2}{1-\rho^2} - \frac{(z_1 - \frac{\mu}{1-\rho})^2}{\frac{2\sigma^2}{1-\rho^2}} - \frac{T-1}{2}log(2\pi) - \frac{T-1}{2}log(\sigma^2) - \sum_{t=2}^{t}\frac{(z_t - \mu - \rho z_{t-1})^2}{2\sigma^2} \tag{4.3.4}$$

where σ^2 denote the variance for ϵ_t. Let $L(\mu, \sigma^2, \rho)$ denote the likelihood function for the function with the unknown parameters.

We can formulate a LR test for independence as

$$Ber_{ind} = LR_{ind} = -2(L(\widehat{\mu}, \widehat{\sigma}^2, 0) - L(\widehat{\mu}, \widehat{\sigma}^2, \widehat{\rho})) \tag{4.3.5}$$

where hats denote the estimated values for the parameters. Under the null hypothesis the LR

test statistic for independence is chi-squared with one degree of freedom since the number of restrictions is one, $Ber_{ind} \sim \chi^2(1)$.

We can also use the LR framework to jointly test for both independence and that the series has mean and variance equal to 0 and 1. The combined statistic is formulated as follows

$$Ber = LR = 2(L(0,1,0) - L(\widehat{\mu}, \widehat{\sigma}^2, \widehat{\rho})). \tag{4.3.6}$$

Under the null hypothesis, the LR test statistic is distributed as chi-squared with three $d.o.f.$ since the number of restrictions is three, $Ber \sim \chi^2(3)$.

4.4 Evaluating Density Forecasts with Likelihood Ratio and Markov Chains

The methods presented by Diebold, Gunther and Tay (1998) and Berkowitz (2001) to evaluate density forecasts require specifying a parametric method and the specification of time dependence. Li and Andersson (2018) presented a method, based on Christofferesen's (1998) method, which can be used to evaluate non-parametric models. The method is constructed in the following three steps: a goodness of fit test, an independence test and a joint test for goodness of fit and independence.

The following is the first part of the evaluation method presented by Li and Andersson (2018), a test for goodness of fit for density forecasts. We can measure the goodness of fit by using a unconditional test statistic LR_{ud} for density forecasts. First step is to consider the ex-post outcome, $Y = (y_1, y_2, \ldots, y_t)$, generated by the distribution $f(y_t)$, and the ex-ante forecasted density $s(y_t)$. Let $[I_0, I_n]$ denote the range of y_t, i.e. $I_0 < y_t < I_n$, then divide $[I_0, I_n]$ into k mutually exclusive states as $[\underbrace{I_0, I_1}_{1}, \ldots, \underbrace{I_{k-1}, I_n}_{k}]$, and let n_i denote the number of

y_i which lie in state i. This is an extension of the interval forecasting evaluation and interval forecasting is a special case where $k = 2$. As the Diebold, Gunther and Tay (1998) method, we evaluate if $s(y_t)$ have the correct description of the unconditional probabilities of future values by testing $s(y_t) = f(y_t)$. Under the null hypothesis, forecasted distribution is equal to the true data generating process (DGP), $s(y_t) = f(y_t)$ and the set of probabilities of y_i being in state i, $N = (n_1, \ldots, n_k)$, follows a multinomial distribution, $multinom(T, p_1, \ldots, p_k)$, with event probability $p_i = \int_{I_{i-1}}^{I_i} s_y(u)du$. The likelihood function under the null hypothesis is

$$L(p) = \frac{T!}{(n_1! \ldots n_k!)} p_1^{n_1} \ldots p_k^{n_k} \tag{4.4.1}$$

where $p_i = \int_{I_{i-1}}^{I_i} s_y(u)du$, and the likelihood under the alternative hypothesis is

$$L(\widehat{p}) = \frac{T!}{(n_1! \ldots n_k!)} \widehat{p}_1^{n_1} \ldots \widehat{p}_k^{n_k} \tag{4.4.2}$$

where $\widehat{p}_i = n_i/T$ is the maximum likelihood estimate of the event probability over the whole parameter space. Therefore can the test statistic for unconditional coverage can be formulated as a standard likelihood ratio test

$$LR_{ud} = -2 \log \frac{L(p)}{L(\widehat{p})} \tag{4.4.3}$$

and under the null hypothesis the statistic is distributed chi-squared with $k - 1$ degrees of freedom,

$$LR_{ud} \sim \chi^2(k-1). \tag{4.4.4}$$

Such as the unconditional coverage test LR_{uc} for interval forecasts, the LR_{ud} can be viewed as a pure goodness of fit test.

The next test for density evaluation is for independence and it is an extension of the test for independence for interval forecast. The independence hypothesis is evaluated against a k state first order Markov chain. Let $\pi_{ij} = P(y_t \in state\ j | y_{t-1} \in state\ i)$ and let $\Pi =$

$$\begin{bmatrix} \pi_{11} & \cdots & \pi_{1k} \\ & \ddots & \\ \vdots & \pi_{i,j} & \vdots \\ & \ddots & \\ \pi_{k1} & \cdots & \pi_{kk} \end{bmatrix}$$ denote the transition matrix for the Markov chain. The likelihood function under the alternative hypothesis is

$$L(\Pi) = (\pi_{11}^{n_{11}} \dots \pi_{1k}^{n_{1k}}) \dots (\pi_{i1}^{n_{i1}} \dots \pi_{ik}^{n_{ik}}) \dots (\pi_{k1}^{n_{k1}} \dots \pi_{kk}^{n_{kk}}) = \prod_{i=1}^{k} \prod_{j=1}^{k} \pi_{ij}^{n_{ij}} \tag{4.4.5}$$

where n_{ij} is the number of events where a state i is followed by a state j and $\widehat{\pi}_{ij} = \frac{n_{ij}}{\sum_{j=1}^{k} n_{ij}}$being the MLE for π_{ij}. Under the null hypothesis of independence past information does not influence the present outcome. Hence, if an outcome y_t is in state j, all the previous outcome y_{t-1} has the same probability being in any of the states. This can be denoted as $\pi_{1j} = \pi_{2j} = \cdots = \pi_{kj} = \pi_{.j}$ and under the null hypothesis we have

$$(\pi_{11}^{n_{11}} \dots \pi_{1k}^{n_{1k}}) \dots (\pi_{i1}^{n_{i1}} \dots \pi_{ik}^{n_{ik}}) \dots (\pi_{k1}^{n_{k1}} \dots \pi_{kk}^{n_{kk}}) = \prod_{j=1}^{k} \pi_{.j}^{n_{ij}} \tag{4.4.6}$$

where $n_{.j} = \sum_{i=1}^{k} n_{ij}$. Since $n_{.j}$ is the number of outcomes that lies in state j and $\pi_{.j}$ is the probability that an outcome lies in state j, the MLE for $\pi_{.j}$ is $\widehat{\pi}_{.j} = n_{.j}/T$ and $n_j = n_{.j}$. Under the null hypothesis the likelihood function is

$$L(\widehat{\Pi}_0) = \prod_{j=1}^{k} (\frac{n_j}{T})^{n_j}, \tag{4.4.7}$$

and the likelihood function under the unrestricted, alternative hypothesis is

$$L(\widehat{\Pi}_1) = \prod_{i=1}^{k} \prod_{j=1}^{k} (\frac{n_{ij}}{\sum_{j=1}^{k} n_{ij}})^{n_{ij}}. \tag{4.4.8}$$

Hence the likelihood ratio test (LRT) statistic for independence is

$$LR_{id} = -2 \log \frac{L(\widehat{\Pi}_0)}{L(\widehat{\Pi}_1)} \sim \chi^2((k-1)^2). \tag{4.4.9}$$

By having a closer look at

$$L(\widehat{\Pi}_0) = \prod_{j=1}^{k} (\frac{n_j}{T})^{n_j} = \prod_{j=1}^{k} (\frac{\sum_{i=1}^{k} n_{ij}}{T})^{\sum_{i=1}^{k} n_{ij}} \tag{4.4.10}$$

and

$$L(\widehat{p}) = \frac{T!}{(n_1! \dots n_k!)} \widehat{p}_1^{n_1} \dots \widehat{p}_k^{n_k} = \frac{T!}{(n_1! \dots n_k!)} (\frac{n_1}{T})^{n_1} \dots (\frac{n_k}{T})^{n_k} \tag{4.4.11}$$

we notice that $L(\widehat{\Pi}_0)$ is proportional to $L(\widehat{p})$, i.e. $L(\widehat{\Pi}_0) \propto L(\widehat{p})$. This fact simplifies the next test, the test for the conditional density.

The next test is to test if the conditional forecasted density distribution based on the past information, $s(y_t)|\Omega_{t-1}$, provides the correct conditional probabilities for events related to future actual outcomes. The test statistic is similar to the LR_{cc} in terms of interval forecasting, and the test is a combination of a goodness of fit test and independence test since we simultaneously test if $s(y_t) = f(y_t)$ and if $\{y_t\}_{t=1}^{T}$ is independent. The test statistic for this test is constructed using the additivity of the LRT statistic (Bera and McKenzie, 1985). The sum of the test statistics that test the unconditional hypothesis and the independence hypothesis separately is the test statistic we use to test the joint hypothesis. We denote the test statistic for the joint test of independence and goodness of fit as $LR_{cd} = LR_{ud} + LR_{id}$. We have that

$$\begin{aligned} LR_{ud} = -2 \log \frac{L(p)}{L(\widehat{p})} &= -2 \log \frac{\frac{T!}{(n_1! \dots n_k!)} p_1^{n_1} \dots p_k^{n_k}}{\frac{T!}{(n_1! \dots n_k!)} \widehat{p}_1^{n_1} \dots \widehat{p}_k^{n_k}} \\ &= -2[log(p_1^{n_1} \dots p_k^{n_k}) - log(\widehat{p}_1^{n_1} \dots \widehat{p}_k^{n_k})], \end{aligned} \tag{4.4.12}$$

$$\begin{aligned} LR_{id} = -2 \log \frac{L(\widehat{\Pi}_0)}{L(\widehat{\Pi}_1)} &= -2 \log \frac{\prod_{j=1}^{k} (\frac{n_j}{T})^{n_j}}{\prod_{i=1}^{k} \prod_{j=1}^{k} (\frac{n_{ij}}{\sum_{i=1}^{k} n_{ij}})^{\sum_{i=1}^{k} n_{ij}}} \\ &= -2[\log \prod_{j=1}^{k} (\frac{n_{\cdot j}}{T})^{n_{\cdot j}} - \log \prod_{i=1}^{k} \prod_{j=1}^{k} (\frac{n_{ij}}{\sum_{i=1}^{k} n_{ij}})^{\sum_{i=1}^{k} n_{ij}}] \end{aligned} \tag{4.4.13}$$

and $\widehat{p}_j = n_j/T$. The $LR_{cd} = LR_{ud} + LR_{id}$ can be simplified as

$$LR_{cd} = -2[\log(p_1^{n_1} \dots p_k^{n_k}) - \log(\prod_{i=1}^{k}\prod_{j=1}^{k} \frac{n_{ij}}{(\sum_{j=1}^{k} n_{ij})^{n_{ij}}})] \sim \chi^2(k(k-1)) \tag{4.4.14}$$

where $p_i = \int_{I_{i-1}}^{I_i} s_y(u)du$. The LR_{cd} can both test for mis-specified density forecast and internal dependence of the data series, compared to LR_{ud} which can only test for biased unconditional forecasted density and ignores the potential internal dependence of $\{y_t\}_{t=1}^{T}$. As a result, LR_{cd} can discover time dependence such as autocorrelation or conditional heteroscedasticity in the forecast errors, instead of only testing the unbiasedness of the forecasted distribution. Under the null hypothesis for the LR_{cd} the forecasted density distribution is equal to the true DGP, $s(y_t) = f(y_t)$, and $\{y_t\}_{t=1}^{T}$ is independent. The test can be applied to evaluate the efficiency of the density forecasts.

The three LR tests in this section can be applied in a natural sequence, in the same manner as we can apply the Chrisofferesen's (1998) tests for interval forecasts. First we jointly test for goodness of fit and independence using the LR_{cd} test. If we accept the null hypothesis we conclude that the forecasting model captures the time dependence in the data and that the distribution in the null hypothesis is the correct distribution. If we reject the null hypothesis we further investigate why by applying the LR_{ud} and the LR_{id} separately. Then we can discover if we rejected the null hypothesis due to lack of independence or if we incorrectly specified the distribution.

Chapter 5

Monte Carlo Simulation

5.1 Introduction to Size and Power of a Test and Monte Carlo Simulation

In this chapter we will use a Monte Carlo study to test the size and power of the test proposed by Berkowitz (2001) and Li and Andersson (2018). To see how well the tests perform we will compare the size and power of the tests in the study. Both Berkowitz (2001) and Li and Andersson (2018) proposed tests for independence and conditional coverage, and we will compare these tests. Li and Andersson (2018) also suggested a test for unconditional coverage, and to see how well this performs we will compare it to the KS test. Shortly explained, the KS test is a non-parametric test to measure the equality of one-dimensional, continuous probability distributions (Chkravarti, Laha and Rot, 1967). It can be used to compare a sample with a specified probability distribution or compare two samples with each other to check if they have the same distribution.

When testing hypothesis in statistics, two types of errors can occur. Type 1 error is when the null hypothesis is rejected when it is true and Type 2 error is when the null hypothesis is

not rejected when it is false (Hogg, Craig and McKean, 2013).

	Null hypothesis is true	Null hypothesis isn't true
Reject null	Type 1 Error	OK
Don't reject null	OK	Type 2 error

Suppose we have a parameter θ in parameter space Θ and we wish to test the hypothesis

$$H_0 : \theta \in \Theta_0 \; vs \; H_A : \theta \in \Theta_A \tag{5.1.1}$$

where Θ_0 and Θ_A denote partitions of the parameter space Θ. First we decide the significance level of a test, denoted by α. The probability of rejecting H_0 depends on the true value of θ and let $\pi(\theta)$ denote the probability of rejecting H_0. Then the significance level of a test is defined as

$$\alpha = \sup_{\theta \in \Theta_0} \pi(\theta) \tag{5.1.2}$$

i.e. α is the largest value of $\pi(\theta)$ when $\theta \in \Theta_0$. The significance level α is an upper bound on the probability of a Type 1 error and it is set before the test, therefore the Type 1 error is controlled by α.

We can use a Monte Carlo study to investigate the size of a test, i.e. the empirical Type 1 error rate. The first step is to generate a data sample under the null hypothesis, $H_0 : \theta \in \Theta_0$, and compute the test statistic. Then record how many times the test is rejected and calculate the proportion. Suppose we replicate a statistical test m times under the null hypothesis and we reject the null hypothesis k times. Then the proportion is $\frac{k}{m}$ is the empirical Type 1 error rate $\widehat{\alpha}$. We want the proportion to approximately be α, $\frac{k}{m} \approx \alpha$, and when m is large the Type 1 error rate will approximate α.

The power of a test is denoted by $\pi(\theta)$ when $\theta \in \Theta_A$. Then we can measure the Type 2 error as $P(Type\ 2\ error) = 1 - \pi(\theta)$ when $\theta \in \Theta_A$. The power of a test depends particular value of $\theta \in \Theta_A$, thus the power will be a function of the true value θ. When α is set we prefer a test with high power, i.e. a test with low probability of a Type 2 error. To evaluate the power of a test the first step is to generate a data sample under the alternative hypothesis, $H_A : \theta \in \Theta_A$, and compute the test statistic. Then record how many times H_0 is rejected at significance level α and compute the proportion. Suppose we replicate a statistical test m times under the alternative hypothesis and we reject the null hypothesis l times. Then the proportion $\frac{l}{m}$ is the empirical power $\widehat{\pi}(\theta)$ and when m is large $\widehat{\pi}(\theta)$ will approximate $\pi(\theta)$. The standard error is defined as $se(\widehat{\pi}(\theta)) = \sqrt{\frac{\widehat{\pi}(\theta)(1-\widehat{\pi}(\theta))}{m}} \leq \sqrt{\frac{0.5}{m}}$.

The main purpose of the empirical size and power is to assess a test or compare different tests. This is especially important in the situations where theoretic Type 1 and Type 2 errors have a non trivial solution or have analytic complications.

5.2 Monte Carlo Simulation for Density Forecast Evaluation Methods

In the Monte Carlo study the null hypothesis is that the forecasted density distributions is equal to the true data generating process and the alternative hypothesis is that they are different;

$$H_0 : s(y_t) = f(y_t) \quad vs \quad H_A : s(y_t) \neq f(y_t). \tag{5.2.1}$$

We let $s(.)$ denote the forecasted distribution and let $f(.)$ denote the true distribution of the DGP.

To illustrate an example of how we can use the Monte Carlo method to measure the size

and power for the density forecasting evaluation methods we will use the LR_{ud}, LR_{id} and LR_{cd} tests. We can, for example, use the *i.i.d.* $N(0,1)$ and *i.i.d.* $t(7)$ distributions to measure the size and the power. First for the size property we assume that $s(y_t)$ for $\{y_t\}_{t=1}^{T}$ is distributed as a independent standard normal, *i.i.d.* $N(0,1)$. To measure the empirical Type 1 error rate, the first step is to generate 1000 density forecasts from H_0 and compute the LR_{ud}, LR_{id} and LR_{cd} test statistics. Then decide on a significance level, for example $\alpha = 0.05$, and calculate the critical value. From the previous section we can find the critical value based on the chi-squared distribution and the k states can be chosen based on Sturges' rule for deciding the ideal bin width when constructing a histogram (Sturges, 1926). The rule is that the number of states is chosen as the integer value of $1 + log_2(T)$ where T is the sample size and the interval length from each state is identical. From this we calculate the proportion of rejections and measure the size of the test, i.e. the empirical Type 1 error rate. For example, if the data sample consists of a 1000 replicates and the significance level is 0.05, we want around 50 rejections.

For measuring the power of the LR_{ud} and LR_{cd} we can continue with the same distributions as above, *i.i.d.* $N(0,1)$ and *i.i.d.* $t(7)$. For example, we can measure the power when the distribution $f(.)$ of the DGP is *i.i.d.* $N(0,1)$ while $s(.)$ is *i.i.d.* $t(7)$. First, generate a sample of density forecasts with the t-distribution and evaluate the hypothesis $H_0 : s(y_t) = f(y_t)$. The next step is to compare the test statistics to the appropriate critical values, calculate the number of rejections and measure the power of the test. E.g. say we generate a sample of 1000 density forecasts under $y_t \sim$ *i.i.d.* $t(7)$ and our hypothesis is that the density should be standard normal. We then calculate the proportion of rejections at a significance level $\alpha = 0.05$. The proportion of rejections is the empirical power of the test and we want a high number of rejections, e.g. if we have a sample of a 1000 replicates and we have 900 rejections, then we have power of 90%.

Since both the *i.i.d.* $t(7)$ and *i.i.d.* $N(0,1)$ are independent processes and LR_{id} is a pure independence test, the illustrated Monte Carlo simulations above will measure the size of this test, not the power. To measure the power of the LR_{id} we have to use a process with dependence and for this purpose we can, for example, simulate density forecasts with $GARCH$ dependence.

5.3 Size and Power Table

For the size and power tables, we have divided the forecasted distributions into two cases:

$$Case\ 1 : y_t \sim i.i.d.\ N(0,1),\ y_t \sim i.i.d.\ t(7) \tag{5.3.1}$$

$$Case\ 2 : y_t = n_t\sqrt{h_t};\ h_t = 0.2 + 0.6y_t^2 + 0.2h_{t-1} \tag{5.3.2}$$

$$n_t \sim i.i.d.\ N(0,1);\ n_t \sim i.i.d.\ t(7). \tag{5.3.3}$$

We will use Case 1 to determine the size of the tests and also the power of the LR_{ud} and LR_{cd} test, and we will use Case 2 to measure the power of the three tests. We simulated samples with Monte Carlo replication of 5000 with sample sizes 250, 500 and 1000. We know from the previous section that a standard error is $se(\widehat{\pi}(\theta)) = \sqrt{\frac{\widehat{\pi}(\theta)(1-\widehat{\pi}(\theta))}{m}}$, and with this we can calculate a 95% CI for the estimated size 5%: $0.05 \pm 1.96\sqrt{\frac{0.05(1-0.05)}{5000}} = (0.0469, 0.0531)$.

Table 5.1: Size of the tests when s(.) = f(.)

DGP	i.i.d. N(0,1)			i.i.d. t(7)		
N	250	500	1000	250	500	1000
LR_{ud}	0.0122	0.0156	0.0170	0.0436	0.0490	0.0482
KS	0.0512	0.0512	0.0484	0.0446	0.0516	0.0474
LR_{id}	0.1124	0.0886	0.0864	0.0676	0.0422	0.0324
Ber_{ind}	0.0540	0.0474	0.0544	0.0450	0.0448	0.0534
LR_{cd}	0.0748	0.0594	0.0636	0.0634	0.0444	0.0328
Ber	0.0054	0.0054	0.0028	0.0460	0.0498	0.0518

In Table 5.1 the forecasted distribution is the same as the true data generating process, i.e. $s(.) = f(.)$, and we start with comparing the sizes of the six tests. When $s(.)$ is $i.i.d.$ $t(7)$ the size of all six tests are unbiased or nearly unbiased, and all six tests perform well.

When the forecasted distribution $s(.)$ is $i.i.d.$ $N(0,1)$, the LR_{id} and the LR_{cd} tests tends to be over biased and have a too large rejection rate, especially for the smaller samples. We see that the size bias decrease as the sample increase, and the rejection rate tends to approach 5% for the LR_{cd} test. The LR_{ud} test is under biased, the rejection rate is low and we see a small increase in the size as the sample sizes increase.

For the KS and Ber_{ind} tests the size is unbiased or nearly unbiased for all the sample sizes and both tests perform well when $s(.)$ is $i.i.d.$ $N(0,1)$. On the other hand, the Ber test perform quite poorly and the size is very under biased. The rejection rate is lower than 1% and increase in sample size has little impact on the size distortion.

Table 5.2: Power of unconditional and conditional tests and size of when s is i.i.d. N(0,1)

DGP	i.i.d. t(7)		
N	250	500	1000
LR_{ud}	0.2090	0.4758	0.8504
KS	0.0630	0.0926	0.1834
LR_{id}	0.0665	0.0466	0.0324
Ber_{ind}	0.0435	0.0468	0.0456
LR_{cd}	0.1040	0.1904	0.3854
Ber	0.0030	0.0068	0.0064

Table 5.3: Power of unconditional and conditional tests and size of when s is i.i.d. t(7)

DGP	i.i.d. N(0,1)		
N	250	500	1000
LR_{ud}	0.2212	0.5882	0.9712
KS	0.0558	0.0736	0.1438
LR_{id}	0.1080	0.1016	0.0816
Ber_{ind}	0.0496	0.0554	0.0510
LR_{cd}	0.1876	0.3206	0.6300
Ber	0.5632	0.9114	0.9986

In the Table 5.2 and 5.3 the forecasted distribution are different from the true DGP, i.e. $s(.) \neq f(.)$. The forecasted distributions are from Case 1. Since both the processes are independent we measure the power of LR_{ud} and LR_{cd} and the size of LR_{id}.

In Table 5.2 the Ber test for conditional coverage has barely any power. In his article, Dowd (2004) shows a deviation from normality of the transformed data and this makes it difficult to detect deviations from normality. The LR_{ud} and the LR_{cd} have better power properties compared to the KS and Ber tests.

In Table 5.3, the test with the highest overall power is the Ber test. For the LR_{ud} and LR_{cd} the power is low when the sample is small and the power increase as the sample increase.

As we can see both Table 5.2 and 5.3, the KS test shows little power. The power increase when the sample increase, but even when the sample size is 1000 the power is still lower than 20% for both forecasted distributions.

For the independence tests, we see in Table 5.3 the Ber_{ind} the size is unbiased or close for all the sample sizes. The LR_{id} is over biased but the size distortion decrease when we increase the sample size. In Table 5.2 both the independence tests are unbiased or nearly unbiased.

Table 5.4: Power of the tests when DGP is from case 2 and n_t is $i.i.d.$ $N(0,1)$

DGP	i.i.d. N(0,1)			i.i.d. t(7)		
N	250	500	1000	250	500	1000
LR_{ud}	0.5590	0.8622	0.9934	0.9156	0.9932	0.9998
KS	0.6514	0.9100	0.9974	0.8018	0.9762	1.0000
LR_{id}	0.8316	0.9800	0.9994	0.8310	0.9830	0.9988
Ber_{ind}	0.2788	0.3336	0.3706	0.2124	0.2428	0.2506
LR_{cd}	0.8826	0.9954	1.0000	0.9812	0.9984	0.9998
Ber	0.1468	0.1934	0.2548	0.8862	0.9572	0.9894

Table 5.5: Power of the tests when DGP is from case 2 and n_t is *i.i.d.* $t(7)$

DGP	i.i.d. N(0,1)			i.i.d. t(7)		
N	250	500	1000	250	500	1000
LR_{ud}	0.9425	0.9983	1.0000	0.7915	0.9616	0.9974
KS	0.2515	0.4303	0.7386	0.2860	0.4360	0.6706
LR_{id}	0.8855	0.9757	0.9942	0.8815	0.9700	0.9930
Ber_{ind}	0.4175	0.4813	0.5556	0.2610	0.2952	0.3200
LR_{cd}	0.9820	1.0000	1.0000	0.9600	0.9970	1.0000
Ber	0.3110	0.4227	0.5864	0.6540	0.7298	0.8260

In Table 5.4 and 5.5 the density forecasts are simulated with a $GARCH$ dependence from case 2 and we will measure the power against incorrect fit and dependence.

In Table 5.4 the three LR tests perform really well. All the tests have high power, even with small samples, and converges to 1 as the sample increase. The KS also shows good power properties, both when the DGP is *i.i.d.* $N(0,1)$ and *i.i.d.* $t(7)$. The power for the KS has a power over 65% when the sample size is 250 and goes towards 1 as the sample size increases. The Ber_{ind} and Ber tests shows little power compared to the others when the DGP is *i.i.d.* $N(0,1)$. Especially the Ber shows very little power and even when the sample size is 1000, the power is less than 30% when we test if the DGP is *i.i.d.* $N(0,1)$.

In Table 5.5 the tests with the highest power properties are the LR tests for unconditional coverage, independence and goodness-of-fit compared to KS, Ber_{ind} and Ber respectively. When the sample size is 500 or more, all the LR tests has a power over 90% and again we see that the power converges to 1 as the sample size increase. The KS test shows little power

when the sample size is small. The power increases as the sample size increase but the power is below 75% in all the cases in Table 5.5. The Ber_{ind} and Ber tests fail to detect dependence in the case where the DGP is standard normal, and neither of the tests show high power properties. When the DGP is $t(7)$ the Ber test show good properties, but the Ber_{ind} still fail to detect dependence.

Both Table 5.4 and Table 5.5 show that the LR tests shows good power properties, also in the cases where the error term has the same distribution has the DGP. The LR tests outperform the KS and Ber tests when it comes to detecting $GARCH$-type dependence in the simulated processes.

5.4 Remarks of Monte Carlo Simulations

Based on Table 5.1 to 5.5 we conclude that the LR tests shows good size and power properties. When we measure the sizes of the tests, the LR, the KS, the Ber_{ind} and the Ber tests all show good properties. When we measure power for independent processes, the LR tests usually outperforms the rest. Especially in the case when the simulated data is normally distributed, the LR_{cd} tests shows good power properties and the Ber test has barely any power. When it comes to detecting $GARCH$ simulated dependence in the processes, the LR tests outperforms the rest and this is the main advantage with this evaluation method. In almost all of the cases with $GARCH$ dependence the LR has higher power properties. In particular, the Ber_{ind} test shows poor power properties when detecting $GARCH$ dependence and the LR_{id} test has very high power, even with small samples. Also, when we test if the DGP is $i.i.d.$ $N(0,1)$, both the Ber and Ber_{ind} fail to detect dependence.

In the next chapter, where we will see how the LR evaluation method functions for eco-

nomical, financial and insurance data. Both the test for interval forecasts by Christoffersen (1998) and for density forecasts by Li and Andersson (2018) can be carried out in a natural sequence in applications. First, we use the jointly test for goodness of fit and independence and test for coverage with well known probability distributions, more specifically normal, t and gamma distribution. If we don't reject the test, we can conclude that we have specified the proper distribution for the data set and the time dependence in the data set has been captured by the specified forecasting model. If we reject this test, we further investigate why by applying the test for goodness of fit and the test for independence to check if rejected due to dependence, incorrectly specified distribution or both.

Chapter 6

Empirical Studies

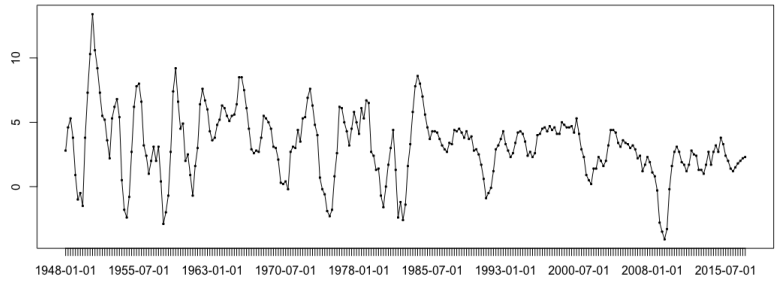
In the following sections we will see how the LR tests for density forecasts perform while being used on empirical economical, financial and insurance data. We will continue with the LR tests based on the Monte Carlo simulation in the previous chapter. In our simulation, the LR tests showed good size and power properties and performed well overall. Also, we have no knowledge of this evaluation method being used on empirical data. Both Diebold, Gunther and Tay (1998) and Rossi (2014) applied the Diebold, Gunther and Tay (1998) evaluation method on empirical data in their papers. More specifically, on Standard & Poor's 500 Index ($S\&P500$) returns and real gross domestic product (GDP) data, respectively. Raaij and Raunig (2002) applied the Berkowitz (2001) evaluation method on financial data, more specifically on daily returns for the Financial Times and Stock Exchange 30 and the $S\&P500$.

We start with a study of the real GDP for the U.S. and continue with a study of $S\&P500$ and the log returns for New York Stock Exchange ($NYSE$) Amex composite index. In the last empirical study we will use insurance data. We will study a time series of compensation amount for fire damage claims in Norway. In the following empirical studies we use a significance level at 5%.

6.1 Real Gross Domestic Product for the U.S.

In the first empirical study we have used a data set for the real GDP in the US. The data is collected on a quarterly basis from 01. January 1948 to 01. July 2017.

Figure 6.1: Real Gross Domestic Product from 01. January 1948 to 01. July 2017



First we are interested to figure out of either the normal- or the t-distribution has the correct conditional coverage for the process. If we reject the hypothesis for conditional coverage we further investigate the sample with tests for independence and unconditional coverage.

We look at the t-distribution first. With the use of the software R, we have estimated the degrees of freedom to be 7. The test statistic for LR_{cd} is 1586.833 and the critical value at a 5% significance level is 74.46832. This means we reject the null hypothesis that the t-distribution provides the correct conditional probabilities for the process. Further we investigate why we rejected the hypothesis by investigating the independence and goodness of fit properties separately. The LR_{id} test statistic is 325.0031 and LR_{ud} test statistic is 1261.83. The critical values are 66.33865 and 14.06714, so we reject the independence and the goodness of fit when we test for coverage with the $t(7)$ distribution.

Next, we check if the normal distribution yields the correct description for the conditional

and unconditional probabilities. Since the independence test is not dependent on the process there is no need to test for independence in this step. We estimate the mean and variance of the sample, 3.193548 and 6.850534, and use the estimates to calculate the test statistics for conditional and conditional coverage. The test statistic for LR_{cd} when we use the normal distribution is 337.2123 and the critical value is 74.46832. Since the conditional coverage test can be viewed as a combination of goodness of fit and independence and we rejected the hypothesis for independence, we are not surprised that we reject the null hypothesis about conditional coverage by the normal distribution also. The next hypothesis we are interested in is if the normal distribution gives an accurate description for the unconditional probabilities, and the LR_{ud} test statistic is 12.20923. The critical value is 14.06713 and since $12.20923 < 14.06713$ we do not reject this hypothesis.

Figure 6.2: Real GDP with fitted normal density

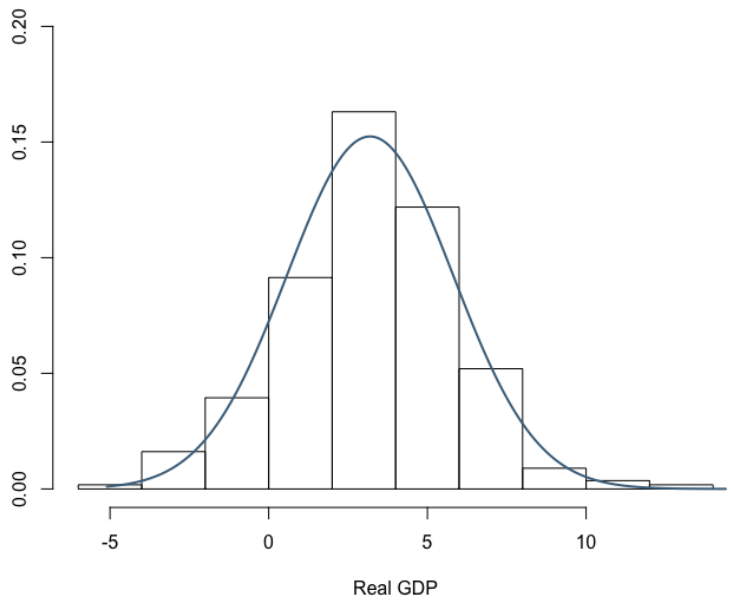
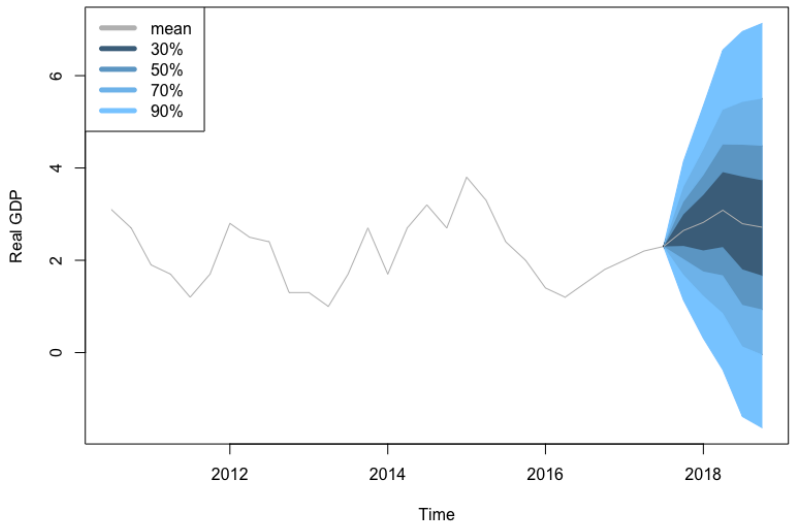


Figure 6.2 is a histogram of the frequencies for the real GDP with an added normal density

with the estimated mean and variance. From a visual perspective it looks like the normal distribution fit the sample well. A possible reason to why we rejected the independence hypothesis and therefore conditional coverage by the normal distribution is that in the data we have included periods when the economy took a hit, for example the financial crisis in 2008.

Figure 6.3: Fan chart for real GDP



Using the statistical software R, we estimate an $ARIMA$ model to model the GDP time series. We have used the function *auto.arima* in R to select the optimal $ARIMA$ model for our GDP time series. The function goes through all the possible models for the time series and chooses the one with the lowest AIC value. The Akaike Information Criterion AIC is measure of the quality of a statistical model and we choose the model with the lowest the AIC value (Wang and Liu, 2006). The AIC value for a model is calculated using the following formula:

$$AIC = 2k - 2ln(L) \tag{6.1.1}$$

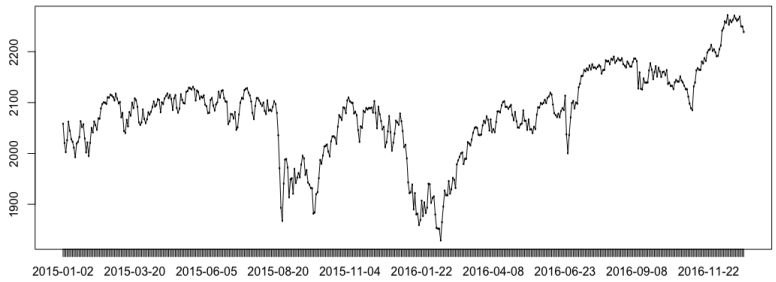
where k is the number of estimated parameters and L denotes the log-likelihood of a model.

Based on the AIC value, the optimal model for the data set is an $ARIMA(3,1,0)\times(0,0,1)_4$ model. With the $ARIMA$ model we can simulate paths for n-steps ahead and get a picture of how the real GDP will develop based on our model. We have chosen to simulate 5000 paths for 5-steps, i.e. 5 quarters ahead. Figure 6.3 shows our predicted development 5 quarters ahead. The means of the process can be looked at as the point forecasts for 5 following quarters. Based on the point forecasts, the model predicts a slight increase in the real GDP the first quarters and then a decrease. The blue fields are the quantiles or the interval forecasts of the process, where the darkest blue field is the 30% interval forecasts and the lightest blue is the 90% interval forecasts. As expected, it fans out further into the forecasting as the uncertainty around the forecasts increase.

6.2 Standard & Poor's 500 Index

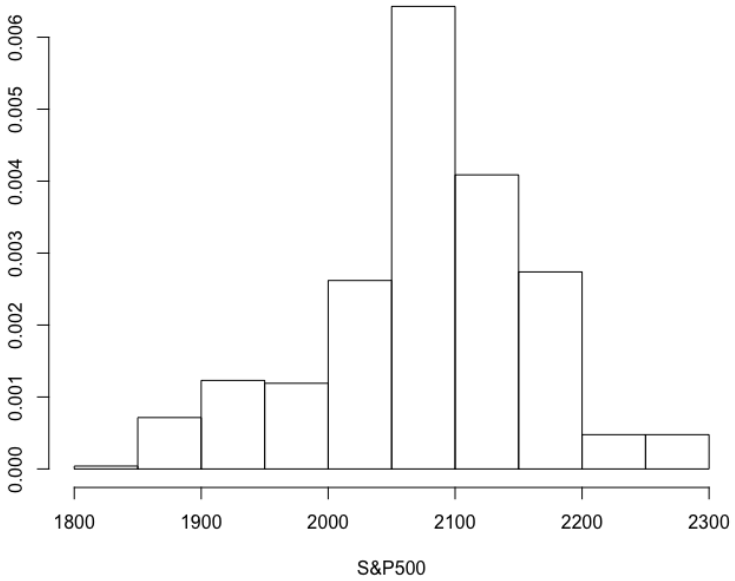
Next we study a time series data-set of the average prices for the $S\&P500$ index. The prices are collected daily in the two year period from 01/01/2015 to 01/01/2017. $S\&P500$ is commonly viewed as a leading indicator of U.S. equities and it is as an index of 500 stocks (Sen and Ma, 2015). The stocks are chosen by S&P Dow Jones Indices and their weights are chosen by a market cap methodology, giving a higher weighting to larger companies (Investopedia, 2018c).

Figure 6.4: S&P500 Price from 1. Jan 2015 to 1. Jan 2017



From Figure 6.5 we see that the data has a bell curve and we will therefore test if the data follows a normal- or a t-distribution, using estimated mean, variance and degrees of freedom.

Figure 6.5: Histogram for S&P500



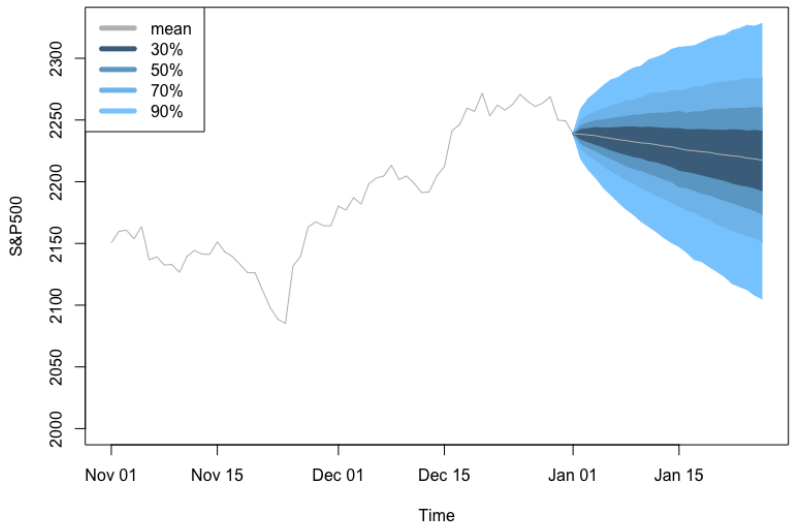
We start with the t-distribution. First we test for conditional coverage using the t-

distribution with estimated degrees of freedom, 6. The value for the LR_{cd} test statistic is 1166.544 and the critical value is 74.46832, so we reject the null hypothesis. Further, we investigate if the t-distribution has the correct unconditional coverage for the data and if the data set is independent. The values for the LR_{ud} and LR_{id} test statistics are 15127.17 and 1166.544, and the critical values are 14.06714 and 66.33865, respectively. This suggest that we reject both hypothesis since both the test statistics are higher than their critical values.

Next we look at the normal distribution with the estimated mean and standard deviation, 2077.86 and 83.18239 respectively. First we check if the normal distribution yields the correct conditional coverage. Since the conditional coverage test is a combination of the independence test and the unconditional coverage test, we technically don't need to apply this test since we already rejected the independence hypothesis. The test statistic LR_{cd} is 1217.46832 and the critical value is 74.46832, which, not surprisingly, implies that we reject the null hypothesis. Next we check if the normal distribution yields the unconditional coverage for the data set. The LR_{ud} test statistic is 50.48608 and the critical value is 14.06714, which means we also reject the hypothesis that the normal hypothesis has the correct unconditional coverage for the $S\&P500$ data set.

Not surprisingly we rejected the independent hypothesis. Financial data is rarely independent, as recognized by Engle (1982) and Christoffersen (1998), and studies have shown that the normal distribution is often insufficient for describing the distribution for financial time series. The standard deviation for financial data is usually heteroskedastic and dependent on the past, and the rejection of the independent hypothesis suggest that the LR test has power on real financial data, not just simulated. The rejection on the independence hypothesis suggest that the $S\&P500$ data can be modeled with an $ARMA - GARCH$ model.

Figure 6.6: Fan chart for S&P500



With the statistical software R we estimate an $ARMA-GARCH$ model for the dataset. The model with the lowest AIC value is an $ARMA(0,2)-GARCH(1,1)$ model. To try to predict how the S&P500 will develop, we simulated 5000 paths of 25 future values. Figure 6.6 shows the predicted development based on the $ARMA-GARCH$ model. The grey line is the mean values or the point forecasts which predicts a decrease for the S&P500. The blue fields are the interval forecasts, where the darkest blue 30% and the lightest blue is the 90%. The fan spreads out and shows a whole spectrum of future possible values.

6.3 Log Returns for New York Stock Exchange Composite Index

As we seen in the previous sections, economical and financial data sets are rarely independent. Interest rates and prices are usually dependent on the previous period(s). One method of

transformation a data set to attempt to turn it into a stationary and/or independent time series is to study the log returns. For a time series $\{Y_t\}_{t=1}^{T}$ the logarithm transformation is defined as

$$log(1 + r_t) = ln(\frac{Y_t}{Y_{t-1}}) \tag{6.3.1}$$

where r_t denotes the return of the time series variable at time t.

When analyzing financial data using log returns are very popular. To see why, we look at some of the properties for log returns (Quantivity, 2011). First, we define a return r_t at time t as

$$r_t = \frac{p_t - p_{t-1}}{p_{t-1}} \tag{6.3.2}$$

where p_t denotes the price at time t. One benefit of using returns, instead of the price, when analyzing is the normalization of the variables, we measure all the variables in a comparable metric. For example, we have a \$1 increase of a stock price. If the stock's original price is \$10 we have a 10% increase and if the stock's original price is \$100 we have a 1% increase. The price has increased by the same amount and we can more clearly compare the impact when we study the returns.

Often in finance one assumption is that prices are distributed log normally, which implies that the $log(1 + r_t)$ is normally distributed since

$$1 + r_t = \frac{p_t}{p_{t-1}} = exp^{log\frac{p_t}{p_{t-1}}}. \tag{6.3.3}$$

This property is one of the main reasons log returns are so popular, since the normal distribution is quite common and popular is statistical analysis. Though, the assumption of log-normality of prices is popular it is not always correct for all stocks, indexes etc.

Another property that explains the popularity of log returns in finance is the time-additivity. We have a sequence of n returns and we want to calculate the compounding return, defined as:

$$(1 + r_1)(1 + r_2) \dots (1 + r_n) = \prod_{t=1}^{n} (1 + r_t) \tag{6.3.4}$$

Since we know from probability theory that a product of normal distributed variables is not normal the above equation isn't very useful. Instead, we use recall that the sum of independent normally distributed r.v. and the logarithmic identity

$$log(1 + r_t) = log(\frac{p_t}{p_{t-1}}) = log(p_t) - log(p_{t-1}). \tag{6.3.5}$$

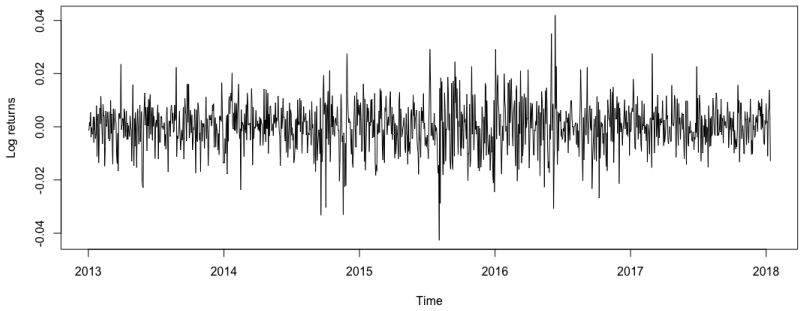
We use this and notice that the compounded log returns can be calculated as

$$\sum_{t=1}^{n} log(1 + r_t) = log(1 + r_1) + log(1 + r_2) + \dots + log(1 + r_n) = log(p_n) - log(p_0). \tag{6.3.6}$$

Then if the prices are log normally distributed and independent, the sum of the natural logarithm of the prices will be normally distributed.

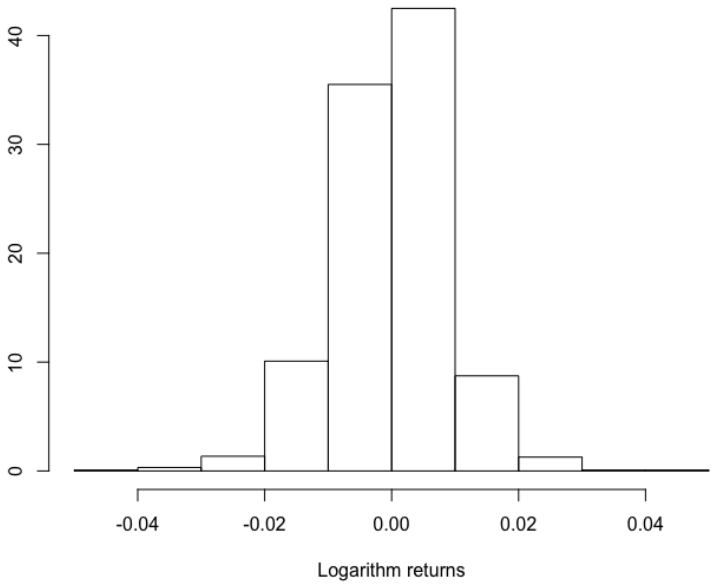
For the log returns study we have chosen a data set of the $NYSE$ Amex composite index. The data set is collected on a daily basis from 28. Jan 2013 to 26. Jan 2018. The $NYSE$ Amex composite index made up of stocks that represent the $NYSE$ Amex equities market (Investopedia, 2018b). The index is a market capitalization-weighted index and the weight of each stock depends on the price of the shares and how many are outstanding.

Figure 6.7: Logarithm returns for NYSE Amex Composite Index



As in the previous section we start by checking if the normal or the t-distribution has the right conditional coverage.

Figure 6.8: Histogram of the logarithm returns for NYSE Amex Composite Index



We start with the t-distribution with estimated degrees of freedom 7. The test statistic

for LR_{cd} is 10845.26 and the critical value is 92.80827, which implies that we reject the null hypothesis that the t-distribution has the correct conditional coverage. Further we investigate why, incorrectly specified distribution, lack of independence or both? The test statistics for these hypothesis are 10764.16 and 81.09181 respectively. The critical value for the goodness of fit test is 15.50731 so we conclude that the t-distribution does not provide the correctly specified density. The critical value corresponding to the independence is 83.67526. Since the test statistic is lower than the critical value we conclude that the sample is independent.

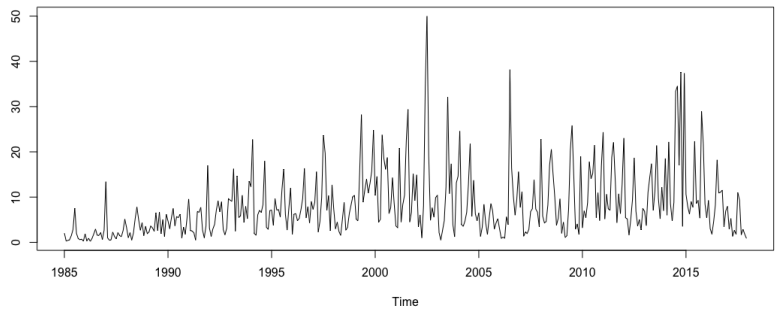
Next we check if the normal distribution has the correct conditional coverage for the estimated probabilities. The test statistic for LR_{cd} is 135.5251 and the critical value is 92.80827, therefore we reject the null hypothesis and we conclude that the normal distribution does not provide the right coverage. Since we already know that the sample is independent and the LR_{cd} test is a combination of goodness of fit and independence, we know that the normal distribution does not provide the right unconditional coverage. The test statistics for goodness of fit LR_{ud} is 54.43333 and the critical value is 15.50731, and as expected we also reject the null hypothesis of unconditional coverage by the normal distribution. Since we reject the hypothesis for coverage by the normal distribution it implies that the prices are not log normal and the theoretical assumption about log normal prices does not always hold for empirical data.

6.4 Compensation amount for fire damage claims in Norway

In this section we will study compensation amounts for fire damage claims caused by electronic equipment in Norway. The amounts are measured monthly from 1985 to 2017 and the value is in 100,000 NOK. Insurance companies can use density forecasting for various purposes, such as predicting the claim amount for different insurances, the rate of traffic accidents resulting

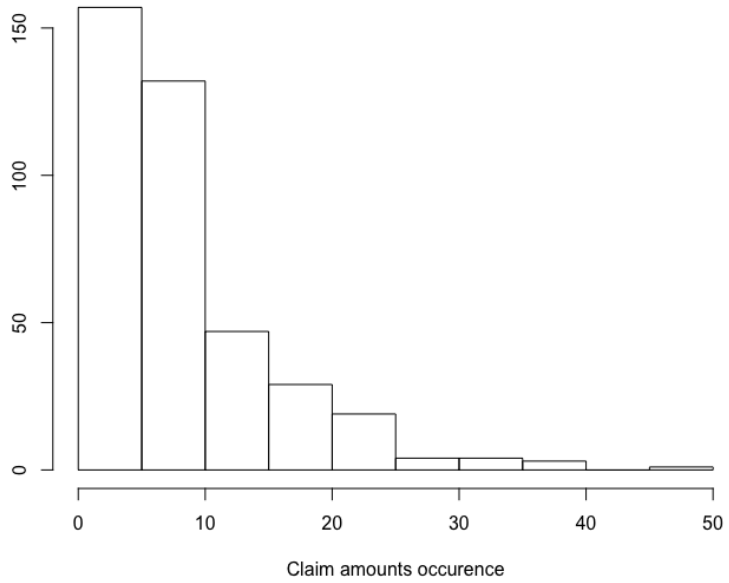
insurance claims etc. With these predictions insurance companies can plan different strategies for different probable outcomes.

Figure 6.9: Compensation amount for fire damage claims in Norway



From a visual perspective of Figure 6.9 the claim amounts look quite random and there is no obvious pattern. We will study this further by checking if the sample is independent. In addition we will also look at known probability distributions to see if any of them contribute the right coverage for the claim amount probabilities.

Figure 6.10: Histogram of claim amounts for fire damage

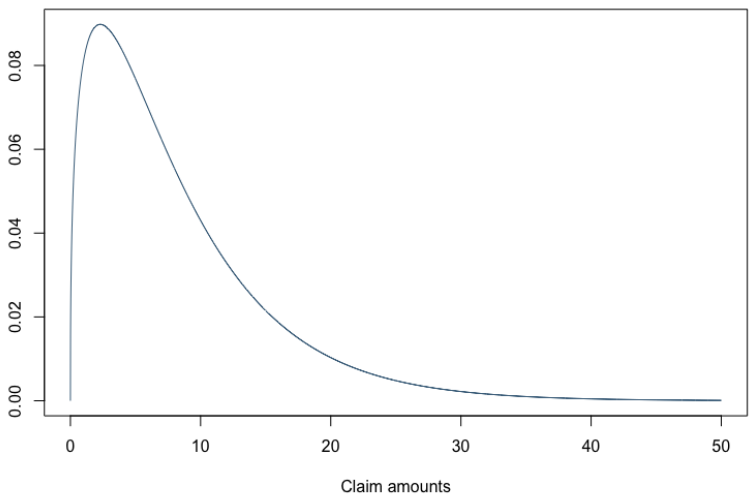


From Figure 6.10 we see that the claims are positively skewed which is quite common in insurance. Incidents that result in smaller claims happen more often, therefore they have a higher probability. The distribution for the claims often have a long tail since an insurance company can experience incidents that have very high claims. Usually these incidents have a low probability. The skewness suggest that the normal- and t-distribution will not supply the correct coverage for the claim probabilities since they are symmetric distributions. In the same procedure as the previous section we will first check if the normal distribution has the correct coverage and if the sample is independent. The LR_{cd} statistic when the null hypothesis is normal distribution has the correct coverage is 98.21342 and the critical value is 58.12404, which implies that we reject the null hypothesis of correct conditional coverage by the normal distribution. Next we investigate the goodness-of-fit and independence separately. The test statistic for goodness-of-fit is 47.77471 and the critical value is 12.19159. We reject the

hypothesis of goodness-of-fit by the normal distribution, which is not surprising by looking at the skewed shape of the histogram. The LR_{id} test statistics is 50.43871 and the corresponding critical value is 50.99846 hence we accept the null hypothesis of independence.

Since the claim amounts are skewed we will check if one of the skewed and continuous distributions has the correct conditional coverage. We have chosen the gamma distribution and with the help of the software R we have estimated the shape to be 1.39 and the rate to be 0.17. The test statistic for conditional coverage is 55.09666 and the critical value is 58.12404, therefore we conclude that the gamma distribution has the correct conditional coverage for the claim amounts. Even though it is not necessary, we also check if the gamma distribution has the correct unconditional coverage. Since we already accepted the null hypotheses of conditional coverage and independence, we should also accept the null hypothesis of unconditional coverage by the gamma distribution. The unconditional test statistic LR_{ud} is 4.657952 and the corresponding critical value is 12.59159. Hence, we accept the null hypothesis of unconditional coverage as expected.

Figure 6.11: One-step ahead forecast for the claim amount



In Figure 6.11 we have made a one-step ahead forecasts for the claims amount based the gamma distribution. The mean value is 8.17 and can be viewed as a point forecast for the claim amount for the next period. The quantiles for 25% and 75% are 3.13099 and 11.23726 respectively. These limits can be viewed as a interval forecast for a 50% level, and there is an estimated 50% probability that the claim amount for the next period should be within these limits.

Chapter 7

Summary and Concluding Remarks

Forecasting future values of variables plays a central and important role when discussing and analyzing time series data. Forecasts are produced and studied daily to attempt to make the best strategies for the future. Since forecasts play an important role in economics and finance, as well as other areas, a natural consequence is that the evaluation of the forecasts play an important role.

Historically, most attention has been focused on construction and evaluating of point forecasts. Point forecasts are simple to understand and compute, and they give a guide to immediate implementation for the forecasts user. Recently, the interest and demand for forecasts that give a description of the uncertainty around the forecast has increased. The interval forecasts can be viewed as the first reaction to the increasing demand for more descriptive forecasts. An interval forecast describes the intervals for the most likely outcomes at different confidence levels and the forecast user can plan different strategies based on the intervals for the forecasted variable. As we have had advances in computer power and statistical methodology, the interest for density forecasts also increased. With technological advances we can produce more advanced density forecasts and don't have to rely on dubious assumptions to

the same degree. Since a density forecast provides a complete description of the uncertainty and since the point- and interval forecasts is merely by-products, the density forecast satisfies all the forecasts users needs. If a user is interested in the point forecasts the mean or the median of the density forecast will suffices, and if a user is interested in the VaR the quantiles of the density forecast is adequate.

A lot of past literature has focused on point forecasts therefore we only gave a brief introduction to evaluating the point forecasts. When evaluating point forecasts, we can look at the forecasting model to see if it satisfies certain desirable properties and we can compare several forecasting models to see if they outperform one another.

Engle (1982) recognized the need for dynamic interval forecasts around the point forecasts, and Christoffersen (1998) identified that if we have dynamic forecasts, we can't evaluate them purely based on unconditional coverage. Since the traditional method for evaluating interval forecasts failed to identify clusters of outliers in the interval forecasts, Christoffersen (1998) developed a method that would identify this issue. The test for conditional coverage is a combination of an independence test and an unconditional coverage test, which identifies if there exist clusters of outliers in the time series.

As the demand for density forecasts increased, we also have seen an increase in the demand for evaluation methods. Evaluation methods for density forecasts has been little explored compared to the point forecast and we chose to focus more on this subject. The three methods we have discussed have all been presented in the interval from 1998 to the present date. The first method we discussed was developed by Diebold, Gunther and Tay (1998). It is based on the PIT of the density forecasts and the uniform distribution. They argue, that if the forecasts densities is correctly specified, then the PIT should be distributed as $i.i.d.$ $U(0,1)$ and there are several tests available to test for uniformity to check this property.

Berkowitz (2001) pointed out that it is difficult to test for uniformity when the sample is small, and proposed a evaluation method based on transformation into the standard normal distribution. If we have specified the right distribution for the density forecasts then the Rosenblatt transformation will transform realizations into $i.i.d.$ $N(0,1)$ variates and use the log-likelihood equation to formulate LR tests for independence and conditional coverage.

Li and Andersson (2018) extended Christoffersen's method for interval forecast evaluation to density forecast evaluation. The three tests they presented are for unconditional coverage, independence and conditional coverage and all three tests are non-parametric. The three tests are based on the likelihood ratio, and the conditional coverage tests is constructed based on the additivity of LRT and is a combination of the independence and unconditional coverage test.

In the Monte Carlo study we compared two methods for density forecasts. Both methods showed good size properties and nearly all the sizes are unbiased for the largest sample size. When we estimate the power of the tests coverage tests using independent processes, the tests proposed by Li and Andersson (2018) perform better than the KS and the test proposed Berkowitz (2001) overall. Especially when the simulated distribution is normal, the Ber test has barely any power while the LR_{cd} test shows increasing power as sample increases. We also measured the power of the tests based on simulating density forecasts with $GARCH$-dependence. The LR tests proposed by Li and Andersson (2018) perform really when the simulated processes has a $GARCH$ dependence, and the main advantage of this evaluation method is its ability to detect $GARCH$-dependence in the processes. The Ber and Ber_{ind} test failed to detect dependence, especially when the DGP is $i.i.d.$ $N(0,1)$.

In the empirical studies used the LR tests to evaluate economical, financial and insurance data. When we studied the real GDP and the $S\&P500$ Index we rejected the conditional

coverage by the normal and the t-distribution hypothesis and the independence hypothesis. This is not surprising since the economical and financial data is usually dependent on the recent realized values. When evaluating the log-returns we did not reject the null hypothesis for independence, but we rejected the hypothesis of goodness of fit for the normal and t-distribution.

The last empirical study was a data set of compensation amounts for fire damage claims caused by electronic equipment in Norway. From a visual perspective, we could see that the claims amounts were quite right skewed which is not unusual for insurance data. We checked first if the normal distribution has the right conditional coverage for the probabilities and we rejected the null. After we checked why by evaluating the independence hypothesis and the goodness-of-fit separately, which led to the conclusion that the sample was independent and the normal distribution did not have the correct unconditional coverage. Since the sample was skewed we checked if the gamma distribution had the correct coverage for the claims, and we accepted the hypothesis. Based on the gamma distribution contributing the correct conditional coverage for the claims, we predicted the claim amount for one-step ahead based on this distribution.

There are several topics that could be interesting to study further. For example to investigate further the evaluation method proposed by Li and Andersson (2018) since we have seen that this evaluation method density forecasts perform well. Since economical and financial data often are dependent on past realization of the target variable and the $GARCH$ model is quite popular, further studies of the density forecast evaluation method with $GARCH$ dependence could be interesting. For example, use a Monte Carlo study to simulate density forecasts with $GARCH$ dependence and measure the size with a testing if true DGP is the same $GARCH$ process. Also measuring the power between different $GARCH$ density forecasts by simulating density forecasts with $GARCH$ dependence and measure the power with

testing if $GARCH$ process with different parameters is the true DGP.

Another more practical study would to look further into the the assumption of log distributed prices for stocks, indexes etc. We can use density forecasting evaluation to investigate this assumption. Our suggestion for further research is to first do a Monte Carlo simulation with the log-normal distribution to measure the size and power of the methods with this distribution. The second step would be testing and evaluating financial data with an evaluation method for density forecasts. Finally, if one don't reject the null hypothesis of log-normal prices, further study the log-returns of this stock or index.

Bibliography

[1] Bao, Y., Lee, T., and Saltoğlu, B. (2007). Comparing Density Forecast Models. *Journal of Forecasting*, 26(3), pp. 203-225.

[2] Bera, A. K. and McKenzie C. T. (1985). Alternative forms and properties of the score test. *Jornal of Applied Statistics*, 13, pp. 13-25.

[3] Berkowitz, J. (2001). Testing Density Forecasts, With Applications to Risk Management. *Journal of Business & Economic Statistics*, 19(4), pp. 465-474.

[4] Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3), pp. 307-327.

[5] Bollerslev, T. and Wooldridge, J. M. (1992). Quasi-Maximum Likelihood Estimation and Inference in Dynamic Models With Time-varying Covariances. *Econometric Reviews*, 11, pp. 143-172.

[6] Brockwell, P. J. and Davis R. A. (1987). *Time Series: Theory and Methods*. 2nd. ed. New York: Springer-Verlag, p. 182.

[7] Brockwell, P. J. and Davis R. A. (2016). *Introduction to Time Series and Forecasting*. 2nd ed. New York: Springer-Verlag, pp. 15-18, 29.

[8] Casella, G. and Berger, R. L. (1990). *Statistical Inference*. 2nd ed. Pacific Grove, Calif: Wadsworth & Brooks, Cole, pp. 388.

[9] Chatfield, C. (1993). Calculating Interval Forecasts . *Journal of Business and Economic Statistics*, 11(2), pp. 121-124.

[10] Chatfield, C. (2000). *Time-Series Forecasting*. Chapman & Hall/CRC, pp. 12-13.

[11] Chakravarti, I. M., Laha R. G. and Roy J. (1967). *Handbook of Methods of Applied Statistics, Volume 1*. New York: John Wiley and Sons, pp. 392-394.

[12] Chiu, C., Lee, M. and Hung, J. (2005). Estimation of Value-at-Risk under jump dynamics and asymmetric information. *Applied Financial Economics*, 15(15), pp. 1095-1106.

[13] Christoffersen, P. F. (1998). Evaluating Interval Forecasts. *International Economic Review*, 39(4), pp. 841-862.

[14] DeGroot, M. H. and Schervish, M. J. (2012). *Probability and statistics*. 4th ed. Boston: Addison-Wesley, p. 626.

[15] Diebold, F. X., Gunther, T.A. and Tay, A. S. (1998). Evaluating Density Forecasts With Applications To Financial Risk Management. *International Economic Review*, 39(4), pp. 863 - 882.

[16] Dowd, K. (2004). A Modified Berkowitz back-test [online] Available at: *https* : *//www.risk.net/risk − management/*1530317/*modified − berkowitz − back − test* [Accessed 9 Mar. 2018].

[17] Engle, R. F. (1982). Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. *Econometrica* 50(4), pp. 987-1002.

[18] Engle, R. F. and Yoo, B. S. (1987). Forecasting And Testing In Co-Integrated Systems. *Journal of Econometrics*, 35(1), pp. 143-159.

[19] Fama, E. F. (1965). Behavior of Stock-Market Prices. *The Journal of Business*, 38(1), pp. 34-105.

[20] Fuller, W. A. and Hasza, D. P. (1981). Properties Of Predictors For Autoregressive Time Series. *Journal of the American Statistical Association*, 76(373), pp. 155-161.

[21] Gneiting, T. (2011). Making And Evaluating Point Forecasts. *Journal of the American Statistical Association*, 106(494), pp. 746-762.

[22] Granger, C. W. J. and Newbold, P. (1986). *Forecasting Economic Time Series*. 2nd ed. New York: Academic Press, p. 149.

[23] Harvey, A. C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge, U.K.: Cambridge University Press, p. 82.

[24] Hogg, R. V. and Craig, A. T. (1965). *Mathematical Statistics*. New York: Macmillan, pp. 258-265.

[25] Hogg, R. V., Craig, A. T. and McKean, J. W. (2013). *Introduction to Mathematical Statistics*. 7th ed. Harlow: Pearson Education UK, p. 241.

[26] Holton, Glyn A. (2014). *Value-at-Risk: Theory and Practice*. 2nd ed. [Online] Available at: *https* : *//www.value − at − risk.net/copyright/* [Accessed 2 Sep. 2017]

[27] Intensity.com, (2017), *Evaluating Interval and Probability Forecasts*. [Online] Available at: *http* : *//intensity.com/wp − content/uploads/*2016/07/$Forecasting_Note_No3.pdf$ [Accessed 4 Sep. 2017].

[28] Investopedia.com, (2018a). *Forecasting*. [online]. Available at: *https* : *//www.investopedia.com/terms/f/forecasting.asp* [Accessed 9 Feb. 2018].

[29] Investopedia.com, (2018b). *NYSE Amex Composite Index*. [Online] Available at: *https* : *www.investopedia.com/terms/n/nyse − amex − composite − index.asp* [Accessed 5 Feb. 2018]

[30] Investopedia.com, (2018c). *Standard & Poor's 500 Index - S&P 500*. [Online] Available at: *https* : *//www.investopedia.com/terms/s/sp*500.*asp* [Accessed 5 Feb. 2018].

[31] Li, Y. and Andersson, J. (2018). A Likelihood Ratio And Markov Chain Based Method To Evaluate Density Forecasting. *Journal of Forecasting* Under minor revision

[32] Medium.com, (2018). *Time Series Analysys for Data IV - ARMA Models*. [Online] Available at: *https* : *//medium.com/auquan/time − series − analysis − for − finance − arma − models −* 21695*e*14*c*999 [Accessed 23. Apr. 2018]

[33] Montgomery, D. C., Jennings, C. L. and Kulahci, M. (2016). *Introduction to time series analysis and forecasting*. 2nd ed. Hoboken, New Jersey: Wiley, pp 2-5.

[34] Quantivity.wordpress.com, (2011). *Why Log Returns*. [Online] Available at: *https* : *//quantivity.wordpress.com/*2011/02/21/*why − log − returns/* [Accessed 8 Feb. 2018].

[35] Raaij, G. d. and Raunig, B. (2002). *Evaluating Risk Models with Likelihood Ratio Tests: Use with Care!* [Online] Available at: *http : //www.fbv.kit.edu/symposium/9th/papers/Rau$_R$aa.pdf* [Accessed 8 Apr. 2018]

[36] Rosenblatt, M.(1952). Remarks on a Multivariate Transformation. *The annals of Mathematical Statistics*, 23, pp. 470-472.

[37] Rossi, B. (2014). Density Forecasts in Economics and Policymaking. *Els Opuscles del CREI*, [Online] pp. 1-31. Available at: *http : //www.crei.cat/wp − content/uploads/opuscles/*140929110100$_E$*NG$_a$ng$_3$7.pdf* [Accessed 6 Sep. 2017]

[38] Sen, R.and Ma, C. (2015). Forecasting Density Function: Application in Finance. *Journal of Mathematical Finance*, 5, pp. 433-447.

[39] Song, J. and Kang, J. (2018). Parameter change tests for ARMA–GARCH models. *Computational Statistics & Data Analysis*, 121, pp. 41-56.

[40] Sturges, H. A. (1926). The Choice Of A Class Interval. *Journal of the American Statistical Association*, 21(153), pp. 65-66.

[41] Tay, A. S. and Wallis, K. F. (2000). Density Forecasting: A Survey. *Journal of Forecasting* 19(4), pp. 235-254.

[42] Wallis, K. F. (2003). Chi-Squared Tests Of Interval And Density Forecasts, And The Bank Of England's Fan Charts. *International Journal of Forecasting* , 19(2), pp. 165-175.

[43] Wang, Y. and Liu, Q. (2006). Comparison of Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) in selection of stock-recruitment relationship. *Fisheries Research*, 77(2), pp. 220-225.

[44] Wei, W. W. S. (1990). *Time Series Analysis: Univariate and Multivariate Methods* 2nd ed. Redwood City, CA: Addison-Wesley, pp. 33, 47, 57, 72, 164-170, 373.

[45] Zarnowitz, V. (1969). The New ASA-NBER Survey of Forecasts by Economic Statisticians. *The American Statistician*, 23(1), pp. 12-16.

Appendices

Appendix A

R-Code

A.1 R-Code for Simulation

```
# AR(1)-LIKELIHOOD
logl=function(th,z)
{
 m=th[1]; s=th[2]; r=th[3]
 n=length(z)
 ll=-0.5*log(2*pi)-0.5*log(s^2/(1-r^2))-(z[1]-m/(1-r))^2/(2*s^2*(1-r^2))
 ll=ll-((n-1)/2)*log(2*pi)-((n-1)/2)*log(s^2)
 for(i in 2:n)
 {
    ll=ll-(z[i]-m-r*z[i-1])^2/(2*s^2)
 }
 nll=-ll
 return(nll)
}

berkowitz=function(z,alternative=alt)
{
   th=c(0,1,0)
   if(alternative=="ar1")
   {
```

```
    meanhat=as.numeric(arima(z, order = c(1,0,0),include.mean=TRUE)$coef[2])
    varhat=as.numeric(arima(z, order = c(1,0,0),include.mean=TRUE)$sigma2)
    ar1hat=as.numeric(arima(z, order = c(1,0,0),include.mean=TRUE)$coef[1])
    llH1=logl(c(meanhat,varhat^0.5,ar1hat),z=z)
  }else
  {
    #ths=c(1,0.1,0.5)
    #opt=nlminb(ths,loglg11,z=z)
    #llH1=opt$objective
    llH1=as.numeric(garchFit(~garch(1,1),include.mean=FALSE,data=z,trace=FALSE)@fit$llh)

  }
  llH0=logl(c(0,1,0),z=z)
  llH0ind=logl(c(mean(z),sd(z),0),z=z)
  LRind=2*(llH0ind-llH1)
  LR=2*(llH0-llH1)
  return(list(LRind=LRind,LR=LR))
}

##################

##I1=1: dependent garch, I2=1: independent iid##
##D1=1:s is t distribution, D2=1:s is normal distribution##
##M1=1: DGP is t distribution, M2=1:DGP is normal distribution##########

Densi=function(D1,D2,T,df,w,a,b,k0,I1,I2,M1,M2,alt,rep){
  LRberind=numeric(0); LRber=numeric(0)
  LRberindtest=numeric(0) ;LRbertest=numeric(0)
  LRuc=numeric(0); LRind=numeric(0); LRcc=numeric(0)
  ks=numeric(0); PIT=numeric(0)
  LRuctest=numeric(0); LRindtest=numeric(0); LRcctest=numeric(0)
  kstest=numeric(0); PITtest=numeric(0)

  k0 <- round(1 + log2(T))
```

```
for (n in 1:rep){
  y=numeric(0)
  y1=numeric(0)
  h1=numeric(0)
  n1=rt(T,df)
  y1[1]=rt(1,df)
  h1[1]=0.1
  for (i in 2:T){
    h1[i]=(w)+a*(y1[i-1])^2+b*h1[i-1]
    y1[i]=n1[i]*(h1[i])^0.5*I1+n1[i]*I2
  }

  y2=numeric(0)
  h2=numeric(0)
  n2=rnorm(T,0,1)
  y2[1]=rnorm(1,0,1)
  h2[1]=0.1
  for (i in 2:T){
    h2[i]=(w)+a*(y2[i-1])^2+b*h2[i-1]
    y2[i]=n2[i]*(h2[i])^0.5*I1+n2[i]*I2
  }
  y=y1*M1+y2*M2 ##DGP##

  pit=numeric(0)
  pit1=rep(0.00001,T)
  xt=numeric(0)
  for (i in 1:T) {
    xt[i]=pt(y[i],df)
  }

  xnorm=numeric(0)
  for (i in 1:T) {
    xnorm[i]=pnorm(y[i],mean(y),sd(y))
  }

  pit=xt*D1+xnorm*D2
```

```
## If pit is 1, replace with 0.999 since qnorm(1) is infinity
for (i in 1:T){
  if(pit[i] == 1.000000e+00){
    pit[i] = 0.999
  }
}

ros=qnorm(pit)

LRberind[n]=as.numeric(berkowitz(ros, alternative=alt)$LRind)
LRber[n]=as.numeric(berkowitz(ros, alternative=alt)$LR)



## Calculate number of rejections
if (LRberind[n]>=qchisq(0.95,1))
{LRberindtest[n]=1}
else
{LRberindtest[n]=0}

if(LRber[n]>=qchisq(0.95,3))
{LRbertest[n]=1}
else
{LRbertest[n]=0}

## Markov Chain Test

low=as.numeric(min(y))    ##lowest value of simulated y##
upp=as.numeric(max(y))    ##highest value of simulated y#

wide=(upp-low)/k0                    #interval width##
divid=seq(low,upp,wide)
divid
divid1=divid ##initial divide point with lowest and highest end value##

x0=numeric()
```

```
x0[1]=count(y<=divid[2])

for (i in 2:(k0)){
  x0[i]=count(divid[i]<y&y<=divid[i+1]+0.00001) ##count the number of
  ##data fall in each interval##
}

nonzero=which(x0!=0) #if the interval contains 0 data set, we do not
#need this interval#

dividnew=numeric() ## new divide point garante non-zero deviding##

for (i in 1:length(nonzero)){
  dividnew[i]=divid[nonzero[i]+1]}

k1=length(dividnew) ##number of new dividing point##

div=numeric()
for (i in 1:(k1-3)) {
  div[i]=dividnew[i+2]
}

k2<- length(div)
x=numeric()
x[1]=count(y <= div[1]) ##number of observations below the lowest divide point#

for (i in 2:(k2)){
  x[i]=count(div[i-1]<y&y<=div[i])
} ##number of observations inside each divide interval##

x[k2+1]=count(y>div[k2]) ##number of observations above the higest
#divide point, means n_i in the paper##

cdf0 <- numeric()
cdfnorm <- numeric()
cdft <- numeric()
```

```
##empirical cumulated density based on different distribution
#assumptions##

for (i in 1:k2) {
  cdfnorm[i]= pnorm(div[i], mean(y), sd(y))
}
for (i in 1:k2) {
  cdft[i]= pt(div[i], df)
}

cdf0 = cdft*D1 + cdfnorm*D2

p0=numeric() ##
p0[1]=cdf0[1]
for (i in 2:(k2)) {
  p0[i]=cdf0[i]-cdf0[i-1]
}

p0[k2+1]=1-cdf0[k2]

LRuc[n] <- -2*sum(x*log(p0*T/x)) ##LRud in the paper##

## We now calculate LRind #
Loguc0 <- sum(x*log(p0))
Loguc1 <- sum(x*log(x/T))

state=numeric(T)

for (t in 1:T){
  if (y[t] <= div[1])
  {state[t]=1}

  for (k in 2:k2) {
    if(div[k-1]<y[t] & y[t]<=div[k])
      state[t]=k
```

```
  }
  if (y[t]>div[k2])
  {state[t]=k2+1}
}

Pi1=matrix(0,(k2+1),(k2+1)) ##caculate matrix pi*T based on observations##
for (t in 2:T) {
  for(i in 1:(k2+1)){
    for (j in 1:(k2+1)){
      if (state[t]==j & state[t-1]==i)
        Pi1[i,j]=Pi1[i,j]+1
    }
  }
}
Pi1 ## counts how many times state i is followed by state j
PI1=matrix(0,k2+1,k2+1) ##get rid of 0 by setting the 0 value in pi to
#0.1, as it will be used in log value#

for(i in 1:(k2+1)){
  for (j in 1:(k2+1)){
    if (Pi1[i,j]==0)
      PI1[i,j]=Pi1[i,j]+0.1
    else
      PI1[i,j]=Pi1[i,j]
  }}

pi=numeric()
for (i in 1:(k2+1)) {
  pi[i]=sum(Pi1[i,])
}

PI=numeric()
for (i in 1:(k2+1)) {
  if (pi[i]==0)
    PI[i]=0.1
```

```
  else
    PI[i]=pi[i]
}

PI2=matrix(0,k2+1,k2+1)

for(i in 1:(k2+1)){
  for (j in 1:(k2+1)){
    PI2[i,j]=PI1[i,j]/PI[i] ## transition matrix
  }
}

LRind[n] <- -2*(Loguc1 - sum(Pi1*log(PI2))) ## LRind ##
LRcc[n] <- LRuc[n]+LRind[n]

### Measuring Empirical Size
if (LRberind[n]>=qchisq(0.95,1))
{LRberindtest[n]=1}
else
{LRberindtest[n]=0}

if(LRber[n]>=qchisq(0.95,3))
{LRbertest[n]=1}
else
{LRbertest[n]=0}

if (LRuc[n]>=qchisq(0.95, k2)){
  LRuctest[n]=1
}
else{LRuctest[n]=0}

if (LRind[n]>=qchisq(0.95, k2^2)){
  LRindtest[n]=1
}
else{LRindtest[n]=0}
```

```
    if(LRcc[n] >= qchisq(0.95, (k2+1)*k2)) {
      LRcctest[n]=1
    }
    else{LRcctest[n]=0}

    ks.norm <- ks.test(y, "pnorm")
    ks.t <- ks.test(y, "pt", df=df)
    p.val <- M2*ks.norm$p.value + M1*ks.t$p.value
    if(p.val >= 0.05){
      kstest[n] = 0
    }
    else{kstest[n]=1}
  }

  ## Results from the simulation, rejection rate
  print("LRbertest"); print(sum(LRbertest)/rep)
  print("LRberindtest");print(sum(LRberindtest)/rep)
  print("LRuc test"); print(sum(LRuctest)/rep)
  print("LRind test"); print(sum(LRindtest)/rep)
  print("LRcc test"); print(sum(LRcctest)/rep)
  print("KS test");print(sum(kstest)/rep)
}
```

A.2 R-Code for Empirical Study

```
empirical <- function(M1, M2){
  LRuc <- numeric(); LRind <- numeric(); LRcc <- numeric()
  h1 <- numeric(); h2 <- numeric(); h3 <- numeric()

  y <- data
  T <- length(data)

  k0 = round(1 + log2(T))
  low=as.numeric(min(y))    ##lowest value of the data##
  upp=as.numeric(max(y))    ##highest value of the data#
```

```
wide=(upp-low)/k0                    #interval width##
divid=seq(low,upp,wide)
divid
divid1=divid ##initial divide point with lowest and highest end value##

x0=numeric()
x0[1]=count(y<=divid[2])

for (i in 2:(k0)){
  x0[i]=count(divid[i]<y&y<=divid[i+1]+0.00001) ##count the number of
  ##data fall in each interval##
}

nonzero=which(x0!=0) #if the interval contains 0 data set, we do not
#need this interval#
dividnew=numeric() ## new divide point guarantee non-zero dividing##

for (i in 1:length(nonzero)){
  dividnew[i]=divid[nonzero[i]+1]}

k1=length(dividnew) ##number of new dividing point##

div=numeric()
for (i in 1:(k1-3)) {
  div[i]=dividnew[i+2]
}

k2 <- length(div)

x=numeric()
x[1]=count(y <= div[1]) ##number of observations below the lowest divide point#

for (i in 2:(k2)){
  x[i]=count(div[i-1]<y&y<=div[i])
} ##number of observations inside each divide interval##
```

```
x[k2+1]=count(y>div[k2]) ##number of observations above the highest
#divide point, means n_i in the paper##

cdf0 <- numeric()
cdfnorm <- numeric()
cdft <- numeric()

##empirical cumulated density based on different distribution
#assumptions##

for (i in 1:k2) {
  cdfnorm[i]= pnorm(div[i], mean(y), sd(y))
}
for (i in 1:k2) {
  cdft[i]= pt(div[i], dof)
}
cdf0 = cdfnorm*M1 + cdft*M2

p0=numeric() ##
p0[1]=cdf0[1]
for (i in 2:(k2)) {
  p0[i]=cdf0[i]-cdf0[i-1]
}

p0[k2+1]=1-cdf0[k2]
LRuc<- -2*sum(x*log(p0*T/x)) ##LRud in the paper##

## We now calculate LRind #
Loguc0 <- sum(x*log(p0))
Loguc1 <- sum(x*log(x/T))

state=numeric(T)

for (t in 1:T){

  if (y[t] <= div[1])
```

```
  {state[t]=1}
  for (k in 2:k2) {
    if(div[k-1]<y[t] & y[t]<=div[k])
      state[t]=k
  }
  if (y[t]>div[k2])
  {state[t]=k2+1}
}

Pi1=matrix(0,(k2+1),(k2+1)) ##caculate matrix pi*T based on observations##


for (t in 2:T) {
  for(i in 1:(k2+1)){
    for (j in 1:(k2+1)){
      if (state[t]==j & state[t-1]==i)
        Pi1[i,j]=Pi1[i,j]+1
    }
  }
}
Pi1 ## counts how many times state i is followed by state j

PI1=matrix(0,k2+1,k2+1) ##get rid of 0 by setting the 0 value in pi to
#0.1, as it will be used in log value#
for(i in 1:(k2+1)){
  for (j in 1:(k2+1)){
    if (Pi1[i,j]==0)
      PI1[i,j]=Pi1[i,j]+0.1
    else
      PI1[i,j]=Pi1[i,j]
  }}

pi=numeric()
for (i in 1:(k2+1)) {
  pi[i]=sum(Pi1[i,])
}
```

```
PI=numeric()
for (i in 1:(k2+1)) {
  if (pi[i]==0)
    PI[i]=0.1

  else
    PI[i]=pi[i]
}

PI2=matrix(0,k2+1,k2+1)
for(i in 1:(k2+1)){
  for (j in 1:(k2+1)){
    PI2[i,j]=PI1[i,j]/PI[i] ## transition matrix
  }}

LRind <- -2*(Loguc1 - sum(Pi1*log(PI2))) ## LRind ##
LRcc <- LRuc+LRind

### Comparing statistics and c.v, see if we accept null
if(LRuc[1] < qchisq(0.95, k2) ) {
  h1[1] <- 0
}
if(LRuc[1] >= qchisq(0.95, k2)){
  h1[1] <- 1
}

if(LRind[1] < qchisq(0.95, (k2)^2)) {
  h2[1]<- 0
}
if(LRind[1] >= qchisq(0.95, (k2)^2)){
  h2[1] <- 1
}
if(LRcc[1] < qchisq(0.95, (k2+1)*k2)) {
  h3[1] <- 0
}
if(LRcc[1] >= qchisq(0.95, (k2+1)*k2)){
```

```
    h3[1] <- 1
  }

  # Results: Test satistics, c.v., reject or not
  print("LRuc + CV")
  print(LRuc)
  print(qchisq(0.95, k2))
  print("LRind + CV")
  print(LRind)
  print(qchisq(0.95, (k2)^2))
  print("LRc + CV")
  print(LRcc)
  print(qchisq(0.95, (k2+1)*k2))
  size.uncon <- sum(h1)
  size.ind <- sum(h2)
  size.con <- sum(h3)
  print(size.uncon*100)
  print(size.ind*100)
  print(size.con*100)
}
```