Active Portfolio Risk in Practice

David Musazai

Masteruppsats i matematisk statistik
Master Thesis in Mathematical Statistics

Masteruppsats 2012:3
Matematisk statistik
Maj 2012

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Mathematical Statistics
Stockholm University
Master Thesis **2012:3**
http://www.math.su.se

Active Portfolio Risk in Practice

David Musazai*

May 2012

Abstract

It is very common that an equity fund is measured against a specified benchmark. To ensure that such a portfolio does not deviate too much from its benchmark, statistical risk measures such as Tracking Error and Value at Risk (VaR) can be applied to the corresponding active portfolio, i.e. the combination of a long position in the portfolio and a short position in the benchmark portfolio. The main purpose of this paper is to study the market risk of active portfolios containing equities. In particular, we wish to investigate how small errors in historical data may affect the VaR calculation. Following an index with a tracker portfolio is never exact. This paper investigates by means of both calculation and simulation (a) what values are at risk compared to the index if different kinds of errors occur, such as missing data, time lags or small random errors, and (b) different modeling choices for the estimation of the Value at Risk (VaR). In the simulations, we use the variance-covariance method for calculating the VaR. The most surprising result of the simulations is that simple linear interpolation gives at least as good results as using the Brownian Bridge approach, no matter what σ is used to calculate the Brownian Bridge. The dominating factor in calculating the VaR is the variance of the index time series. Simulations and calculations have shown that the effects of missing data and time lags are roughly proportional to the prevailing volatility.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: davidmusazai@gmail.com. Supervisor: Joanna Tyrcha.

Preface

I would like to thank my supervisors Joanna Tyrcha at Department of Mathematics, Stockholm University and Jon Blomqvist at Swedbank Robur for their support and guidlines.

A great thanks to Håkan Andersson, who gave me the idea for this thesis and provided me with background knowledge on this topic.

Contents

1 Introduction

According to a Swedish engineering proverb, to know is to measure. From 1987 and onwards, both the European and the American financial systems went through a number of crises, different in conception but common in risk taking beyond capacity. These events brought forward the need to measure risk quantitatively, in order to both aid the decision making in financial institutions, and help regulatory authorities to prevent system failures. While it is an ongoing process, the practitioners in banks, financial institutions and regulatory bodies alike, must use the commonly accepted methods of the day to measure and to an extent manage financial risk.

The first widespread and also the best known of quantitative risk measures is Value at Risk (VaR), made popular by Dennis Weatherstone[1] through his 4:15 report [18]. JP Morgan has even published their methodology, which subsequently won industry-wide acceptance. It was therefore a logical step to create legislation and regulation, which built on the same ground. Legislation is nationwide, while business is international. This insight triggered first the Capital Adequacy Framework, more known under the name Basel accord, and then the New Capital Adequacy Framework, i.e. Basel II [31] (and later on, the Basel III) agreements, to be implemented in national legislation according to agreed schedules. In the course of implementing the Basel agreements I and II, VaR became the *de facto* standard for estimating market risk [10].

1.1 Background

Mutual fund management in the European Union is regulated by the UCITS III ('Undertakings For The Collective Investment Of Transferable Securities') rules [22]. These rules in turn refer to the risk management system of banks, which confirms to the Basel rules. We shall therefore limit us to use VaR for the actual risk calculations.

The risks we are investigating are specific to the active portfolio (see section 2.1), in the sense that they arise in the process of constructing, updating and evaluating the active portfolio. We also limit ourselves to market risks.

1.1.1 Market risk

Market risk [19] is the risk which stems from movements in financial market variables. In general, it includes interest rates, exchange rates, commodity prices, equity prices, etc., but in our case, we will restrict ourselves to equity prices, or more specifically, to a variable constructed from equity prices, i.e.

[1]Dennis Weatherstone was a CEO at JP Morgan during nineties.

an index. Risk is another word for potential financial loss. Market risk in this context means potential financial loss due to positions in financial instruments.

The financial instruments in which we will have positions are stocks and an index. The risks we will investigate are those which are coming from missing stock prices or index values, time lags between price changes and their use in choosing trades, and small random errors in the data underlying buy and sell decisions. These will be investigated one by one. For the bank it is only one step in the process of estimating overall risk. Since the chosen risk measure is not sub-additive it is not possible to estimate risks factor by factor and then take their sum for an overall risk [18], the lesson to learn from such an investigation is the behavior and relative importance of these risk factors, more than just the numerical value from the actual calculations.

1.1.2 Fund management

Fund management is the management of various securities such as shares, derivatives or bonds and other assets in order to meet specified investment goals. These goals are usually given in relation to a specific benchmark. This benchmark in case of a stock portfolio is usually also mainly composed of stocks. The most often used benchmarks are stock indices, essentially imaginary stock portfolios where the portfolio weights might be periodically adjusted to reflect changes in the value of the company in question [32]. How these weights are calculated, and whether they take dividends into account, is not relevant for our investigation. The important thing is that methodologies to calculate these indices are standardized and published. The stock prices that are the basis of the calculations are taken at a specific time on a specific exchange, often the last trade of the day. The relative proportions of the constituent securities are calculated at specified intervals, and they are changing due to the change of stock prices, the number of outstanding stocks, and other factors. They are not usually resulting in proportions which would prompt selling or buying posts of shares. Shares of a specific company may enter and leave the index.

A tracker portfolio is a portfolio that tries to mimic its benchmark. Ideally, it should change its value in the exact proportion its benchmark does (see e.g. [11]). If we invest in the same securities in the exact same proportions the figure in the benchmark, and each time the benchmark composition changes, we succeed to change our investment in the same proportion and at the exact same prices which were used in the benchmark calculations then the tracker portfolio follows the benchmark performance to the letter. Due to the factors mentioned above, maintaining a tracking portfolio is a qualified task, which never succeeds to perfection.

An obvious case when such a portfolio is needed is when an index fund is sold by a financial institute. Tracker portfolios are also needed when derivative products based on indices are sold [3].

1.2 Problems with tracking

Small errors creep in, because it is seldom possible to buy and sell exactly at the prices the benchmark calculations use, and even the proportions are not exactly the same. We would of course like to know how small these errors are, or rather, what their compound magnitude is likely to be – in short, we would like to estimate error inherent in the procedure, and investigate the effect of different sources of errors. These risks sometimes tend to be overestimated [25].

Some of these errors are inherent to the procedure, and we shall not investigate them in detail.

Notably,

- The exact proportion in the index cannot always be followed – only whole posts are bought and sold, certainly not fractions of shares.
- When a company enters or leaves the index, all the index followers should buy or sell in unison, at the closing price of that day, which is practically impossible.
- Exactly replicating the index requires transactions (buying and selling the securities) which in turn implies transaction costs. [24] estimates these at 0.5 % per transaction. High transaction frequency can significantly add to the tracking cost.

In the literature on the tracker portfolio, and its risk measures, the securities in the tracker portfolio are often a fixed subset of the securities in the index [2], [11], [7]. Understandably, managing a portfolio with all 500 companies in S&P 500 is quite tedious; a smaller index like the OMXS30 consisting only of 30 stocks is more manageable. With the advent of robot trading it became possible to track even the indices with hundreds of securities in them with portfolios consisting all those securities, and funds employing this strategy pressed the prices on index funds [8], [27]. Of this reason also, we are here mainly interested in other sources of tracking error than using a subset of the benchmark.

1.3 Importance of the research

The Basel II framework places heavy emphasize on the concept of VaR. Deposit taking institutions are obliged to report VaR figures concerning their market risk daily to the regulatory authorities [17], [1]. Under the Advanced Measurement Approach (AMA), Basel II even prescribes that there can be certain risk measures generated by the bank's internal operational risk measurement system. These in turn are commonly realized using the loss distribution [9]. In any case, the financial institutions have to possess a regulatory capital in proportion to their reported VaR. If the daily losses exceed the daily VaR more than a certain number of times during a 60 days period, the institution has to pay fines to the regulatory authority. To estimate the parameters in a parametric VaR model on historic samples alone has its drawbacks, as demonstrated by [10]. Instead, we can investigate the effect of various factors on the VaR, and draw our conclusions based on their possible or likely values. This is the approach chosen for this paper.

The remainder of this paper is organized as follows. First, we introduce the concept of active[2] portfolio and discuss what kinds of tracking errors shall be investigated in this paper, and which methods will be used to model them. Then we describe the mathematics to model the errors, followed by the mathematics to measure the errors. Following the mathematical models there is a short discussion about the data we use, including a short investigation of Swedish stock market volatilities.

After these preparations, we describe the results of our simulations of these error sources. Both missing data and time lags are simulated in a number of ways: interpolating and extrapolating, linearly or using the standard (Black-Scholes) model, and varying the volatility within its historical bounds. After simulating to investigate the effects of missing data, time lags and small errors on the data series for one selected year of OMXS30 data, we are ready to draw conclusions and provide a simplified description of these effects. Finally, we suggest further studies to both refine our results and make them more robust.

2 Theoretical Framework

2.1 Active portfolio

To gauge the tracker portfolio against its index, the concept of active portfolio is used. A combination of a long position in the tracker portfolio and a short position in the benchmark portfolio is called the active portfolio in

[2] Also called differential portfolio, see e.g. [21].

this paper. To measure the deviation of the tracker portfolio from its benchmark, it is customary to apply risk measures to the active portfolio. The simplest risk measure is the Tracking Error [2], which is defined as

$$E = \frac{\left[\sum_{t \epsilon S} |r_t - R_t|^2\right]^{\frac{1}{2}}}{T}, \tag{2.1}$$

where r_t is the return of the tracking portfolio during a time period, R_t is the return of the index, S is the set of time periods, and T is the number of time periods in S. Tracking Error does not take the drift (of the active portfolio in our case) into account, but the drift is usually much smaller than the volatility, so that is only a minor problem. The variance of the active portfolio is

$$s^2 = \frac{\sum_{t \epsilon S} |r_t - R_t - \mu_{r-R}|^2}{T-1}, \qquad \text{where} \qquad \mu_{r-R} = \frac{\sum_{t \epsilon S} (r_t - R_t)}{T}, \tag{2.2}$$

a value that takes even the drift into account, but it is not a meaningful value for non-mathematicians, and does not fit into an overall system of risk management, being specific to just the active portfolio. VaR eliminates both these shortcomings, providing a threshold value for a certain risk level. The risk measure we shall apply in this paper will therefore be VaR. For its definition, see section 3.2.1. VaR and other risk measures are discussed more in detail in e.g. [23] and [28].

To estimate the aforementioned measures, we have to evaluate historical data: we need a time series for the active portfolio, i.e. a difference between a time series for the tracking portfolio, say PF, and a time series for the benchmark, say BM.

In the remaining of the paper, if not stated otherwise, Y_t will denote the value of the benchmark at time t, $\hat{Y}_t$ will denote the value of the tracking portfolio at time t, V_t will denote the value of the active portfolio at time t.

To show what relation the VaR values of the index respective the active portfolio have, we include here their value for the data we were given. In this paper we assume that the value of the portfolio is equal to one unit of currency, e.g. 1 SEK, 1 USD or 1 EUR. Using the variance-covariance method, the VaR (99% confidence level) of the index was 0.029, while the VaR of the active portfolio was 0.0015. Other methods could be used to calculate VaR. We have used the parametric method (fitting a t-distribution to the data and using the parameters from it to calculate VaR) which gave

0.0348, and the empirical method (counting how many times lost the index more than a certain value) which gave 0.0312. The details of these methods are beyond the scope of the present paper.

The active portfolio is very sensitive to even small errors in the tracking portfolio. Two[3] of these were briefly mentioned in section 1.2, but will not be further discussed. Other sources of errors, which will be investigated in detail, are described below.

Noise Small errors in the data like rounding errors can be simulated by reversing the rounding process. The value 100 can be anything from 99,5 to 100,4 – we just pick a random decimal value from a uniform distribution to each data item in PF. BM can be taken as exact as published. Another, and more general way of dealing with small errors, if rounding errors are not the dominating type of them, is to form the tracker portfolio with help of an error term $\hat{Y}_t = Y_t e^{\varepsilon Z_t}$, where Z_t are i.i.d. standard normal random variables, and ε is a small positive number. The use of this latter procedure to estimate the impact of small errors is motivated by the fact that not being able to buy and sell at the exact closing prices widely outpaces the effects of rounding errors.

Time lags between the two time series arise if they are traded, or based on securities traded, on different exchanges. Usually only end-of-day prices are stored in the available data bases, but exchange rates are published with another time perspective. For example, the Swedish central bank, Riksbanken publishes the exchange rates each day 12:15. The Swedish banks calculate a rate as the mean of buy and sell rates 9:30, the Stockholm stock exchange compiles a rate from those figures 10:05, and that will be the rate published by Riksbanken [33]. This makes the effect of time lag difficult to measure very precisely.

- For simplicity, we assume a time lag between 0 and 1 for the index investigated. The interpretation of this number is not important to our investigation. What is important and makes the approach feasible is that the price of the same security at the same time on two different exchanges follows each other closely enough to motivate the replacement of a price on a foreign exchange with the price of the security on the home exchange at the same time (if traded at the same time).

Missing data in the PF time series is the error source which is investigated at the greatest detail in this paper. If trading in a share is suspended,

[3]Errors coming from companies leaving or entering the index and errors due to transaction costs.

there is no price published that day, but the index calculation usually takes the last day's average value for the daily price [32], depending on the index methodology. This can be taken into account in the tracking portfolio. If the tracker portfolio by some reason does not include the share, there is no data in the PF for it − but the index takes the share into account. Missing data can even stem from simple clerical error.

- If we simply supply the BM value for the missing PF values, we may underestimate the risks. If there are only a few missing values, this can still provide a rather realistic estimation. However, we usually prefer a slight over-estimation to a slight under-estimation when we talk about risks.
- We can also supply estimation for the missing value
  - It can be done by means of a linear interpolation.
  - We assume that the asset price develops according to the standard model[4]

    $$Y_T = Y_t e^{\mu(T-t)+\sigma(W_T - W_t)} \tag{2.3}$$

    where Y_t is the asset price at time t, W_t is a standard Wiener process, σ is a nonnegative real quantity called the volatility, and μ is a real number called the drift. Knowing σ and μ we can use the Brownian Bridge (see section 3.3.2) to construct an estimation for the asset price at time t, $u < t < T$ if we know the prices at T and u.
  - We can even use linear extrapolation, or we can extrapolate with help of the Brownian motion.

In the rest of this paper, we will mainly concentrate on linear and Brownian interpolations.

2.2 Modeling the errors

It would be too complicated to model the price development of every security and introduce errors in them. Instead, we recourse to introducing errors in the value of the tracking portfolio in a way that mimics the impact of errors mentioned above, namely missing data, time lags and (white) noise.

As a starting point, assume that everything went as good as theoretically (but not practically) possible, and the tracking portfolio has exactly the same value all the time as the benchmark. Then let us introduce various sorts of errors by changing the value of the tracking portfolio, and investigate what the effect is.

[4] By standard model we mean the Black-Scholes model or geometric Brownian motion [15].

2.2.1 Modeling missing data

Data can be missing, as mentioned, for several reasons: simple clerical error, suspension of trade for a stock or for the whole trade session if certain conditions are met, different holidays in different countries, etc. We model all of them the same way, disregarding the reasons for them. Instead of constructing a tracking portfolio with all 30 stocks of the OMXS30 index, with some data missing for specific stocks at certain days, we just take one asset: the index and regard its value as missing at certain days.

There are then two questions to answer: how to generate missing dates and what value to assume on those dates.

First, the question of what the missing dates are. We can regard missing data tomorrow as independent of whether data is missing today, or not. Our choice is to suppose a weak dependence on the day before, and only on the day before, i.e. we treat days with missing data as Markov chains, which is described in more detail in section 3.1. Once we know when data is missing, we can estimate it by linear interpolation or extrapolation, by means of Brownian Bridge as a more sophisticated approach, or using the standard model to extrapolate. These technics will further be described in section 3.

2.2.2 Modeling time lags

As mentioned above, we suppose that there is the same time lag each day. It is expressed as a fraction of a day, e.g. 0.5 day.

2.2.3 Modeling noise

Noise will be modeled by setting $\hat{Y}_t = Y_t e^{\varepsilon Z_t}$, where Z_t are i.i.d. standard normal random variables, ε is a suitable small positive number, and for all $t \neq u$ the random variables Z_t and Z_u are independent; $e^{\varepsilon Z_t}$ is called the error term.

3 Mathematical Models

We need mathematical models to realize the modeling specified in the previous section. Choosing the mathematical models for generating missing values has two dimensions. One is simplicity: shall we use the simplest possible approach, i.e. linear inter- or extrapolation, or a more sophisticated approach, based on the standard model as described at the end of section 2.1? The other is between interpolation and extrapolation: shall we generate the missing data for today from an interpolation of yesterdays and tomorrows data (possible only for a historical data set), or shall we pretend to be part of the bank's daily VaR generating process, and extrapolate from data up until yesterday? To be on the safe side, we made models for all four possibilities.

Choosing models for time lags leads to similar questions. We shall compare the efficiency of different choices for it.

Finally, error can be modeled by choosing its distribution.

We need also a mathematical model for quantifying the risk, i.e. a model to calculate VaR. Observe that the concept of VaR in itself is not constructive [18], [31]: it does neither specify the distribution family nor the sampling and estimation process. We have to choose those.

3.1 Markov chain

A Markov chain [26] is a stochastic process $\{\xi_i, i = 1, 2, ...\}$ that takes on a finite or countable number of possible values, usually denoted by the set of nonnegative integers $\{0, 1, 2, ...\}$. If $\xi_i = j$, then the process is said to be in state j. We suppose that whenever the process is in state j, there is a fixed probability $p_{jk}, 0 \le p_{jk} \le 1$, that it will be next in state k independently of in which states the process was before.

Since the value p_{jk} represents the probability that the process will make the transition from j to k and it has to make the transition to some state, it holds that $\sum_k p_{jk} = 1$.

A Markov chain is usually described by its transition matrix 3.1

$$\mathbf{P} = \begin{pmatrix} p_{00} & p_{01} & \cdots \\ p_{10} & p_{11} & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix} \tag{3.1}$$

The simplest non-trivial Markov chain (i.e. the simplest with more than one state) is the Markov chain with two possible states. This will be our tool to find the days when data is supposed to be missing.

There are two states (missing and not missing, 0 and 1), and therefore a 2x2 transition matrix 3.2 is needed.

$$\mathbf{P} = \begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix} \tag{3.2}$$

The transition probabilities p_{00} (missing data yesterday; also missing data today) and p_{11} (data available yesterday and also available today) can be arbitrarily chosen, and they determine the rest of the matrix. These will be the probabilities to vary when we want to better understand the effect of missing values.

Given the transition matrix 3.2 (or, equivalently, the two probabilities mentioned above), and an initial state (i.e. if data is missing on the very first day or not), it is easy to generate a realisation of the Markov chain. Suppose we are in state j. MATLAB provides us with a uniformly distributed random number between 0 and 1, and if this number is smaller than p_{j0}, then the next state will be 0; otherwise, it will be 1. This can be done starting from day 1 to the last day in the series.

Once we have the series of states, i.e. we know which days we have missing data, all that remains is to provide values for the missing data. Clearly we have to interpolate (or extrapolate) from known data, the question is how to make this interpolation. Linear interpolation, and (based on the standard Black-Scholes model for price development) Brownian Bridge are the obvious choices. This will be discussed in greater detail in section 3.3.

3.2 Risk measures

As mentioned in section 1, VaR is the most popular and widespread risk measure. It tells basically that with a certain probability, the loss will not be greater than the measure tells. It does not say how big the actual loss is likely to be, in the event a loss occurs; for that purpose, there is the concept of Expected Shortfall. For continuous distributions, Expected Shortfall is the expected value of loss incurred when the VaR loss limit – threshold is exceeded. This concept may be closer to what a layman would call VaR, but the naming convention has been different. In any case, the actual risk measurement process in the actual bank decides what measure to use. Although VaR is the most widespread implementation of Basel II requirements, it is still a choice and not a regulatory or mathematical requirement.

3.2.1 Variance-covariance method for VaR calculations

Let V be the value of the portfolio, and let R_V denote the return of the portfolio during a period of length ∂t, and let $0 < \xi < 1$ be a given number. The portfolio's VaR – VaR is such a number that $P(R_V < -\text{VaR}) = \xi$ [15].

An alternative definition is that for a confidence level α the VaR at this level is the smallest number l such that the probability that the loss $L = -R_V$ exceeds l is no larger than $(1-\alpha)$ [23]. In other words, it is a quantile of the loss distribution. Formally, using the notation in the present paragraph, we can define it as

$$\mathrm{VaR}_\alpha = \inf\{l\epsilon\Re : P(L > l) \leq 1-\alpha\} = \inf\{l\epsilon\Re : F_L(l) \geq \alpha\}, \tag{3.3}$$

where F_L is the cumulative distribution function for losses [23].

To define expected shortfall we need to define the quantile function first. Let F be a cumulative probability distribution function. For $\alpha\epsilon(0,1)$ the α quantile of F is defined as

$$q_\alpha(F) = \inf\{x\epsilon\Re : F(x) \geq \alpha\}. \tag{3.4}$$

We can then define expected shortfall at confidence level α, denoted as ES_α as

$$\mathrm{ES}_\alpha = \frac{1}{1-\alpha}\int_\alpha^1 q_u(F_L)\mathrm{d}u. \tag{3.5}$$

From 3.3, 3.4 and 3.5 follows immediately that

$$\mathrm{ES}_\alpha = \frac{1}{1-\alpha}\int_\alpha^1 \mathrm{VaR}_\alpha(L)\mathrm{d}u. \tag{3.6}$$

Since VaR_α as the function of α is obviously monotone non-decreasing, $\mathrm{ES}_\alpha \geq \mathrm{VaR}_\alpha$. Both expected shortfall and VaR answers the question of what we can expect in the worst case. VaR answers that the loss in those cases will be *at least* this much. Expected shortfall answers that the loss will be *at average* this much. We shall not estimate expected shortfall in this paper, we shall concentrate on VaR instead.

Given the returns of the active portfolio, we can either take the upper α quantile of the empirical loss distribution, or fit a parametric distribution and calculate VaR from the distribution parameters. There are several methods to fit a parametric distribution to a given set of data. One of the most popular ones is the variance-covariance method. Its application to stock returns is described e.g. in [23]. Here is a practical summary of what is actually done.

We shall use logarithmic returns instead of returns, which is a good approximation and provides a result which is easier to interpret.

Let assume that the logarithmic returns of the components of the active portfolio have a multivariate normal distribution, i.e. $X_{t+1} \sim N_d(\underline{\mu}t, \Sigma)$, where $\underline{\mu}$ is the drift vector (for the active portfolio), and Σ is the variance-covariance matrix. The linearized loss operator then takes the form

$$l^{\Delta}_{[t]}(x) = -V_t\omega_t^{'}x, \tag{3.7}$$

where V_t is the portfolio value at time t, and ω_t is the weight vector. From the general rules for normal distributions, follows that the loss distribution [23] for the active portfolio, which is the weighted sum of the benchmark and the tracker portfolio, is

$$L^{\Delta}_{t+1} = l^{\Delta}_{[t]}(X_{t+1}) \sim N(-\omega_t' \underline{\mu}, \omega_t' \Sigma \omega_t). \tag{3.8}$$

The weight vector ω is always [1 -1] since $V = \hat{Y} - Y$.

To use formula 3.8, we need the values of the mean vector (a 1x2 row vector) and the variance-covariance matrix (a 2x2 matrix). The components of the mean vector are the mean values for the respective log returns of time series (tracker portfolio and benchmark), while the variance-covariance matrix is that of the log returns of these two time series.

From here the VaR is easily calculated as the appropriate quantile. Usually, only the variance of this distribution is relevant, since it gives at least an order of magnitude greater contribution to the VaR value than the mean [15]

$$\mathrm{VaR}_{\alpha} = z_{1-\alpha} \sigma \sqrt{\partial t} - \mu \partial t, \tag{3.9}$$

where $z_{1-\alpha}$ is the $1-\alpha$ quantile of the cumulative normal distribution and σ is the appropriate standard deviation.

The variance-covariance method for calculating the VaR for a portfoliio of securities consists of the following steps:

1. Assume that the logarithmic returns have the distribution $X_t \sim N_d(\underline{\mu}, \Sigma)$ and estimate $\underline{\mu}$ and Σ from the available data.
2. Set $\sigma = \sqrt{\omega_t' \Sigma \omega_t}$ and $\mu = -\omega' \underline{\mu}$.
3. Use formula 3.9 to calculate VaR.

We have calculated the variance-covariance matrix, and the mean, for a daily data, so day is the unit of time measurement, which sets ∂t to 1. To put it into one equation calculating daily risk from daily data, we have

$$\mathrm{VaR}_{\alpha} = z_{1-\alpha} \sqrt{\omega_t' \Sigma \omega_t} - \omega' \underline{\mu}, \tag{3.10}$$

In the literature, this approach[5] is called the Relative VaR, see [21]. The meaning of this measure is as follows: if we try to match our short index holding worth \$1 with a long tracking portfolio also worth \$1, then the VaR for this combination is the calculated relative VaR.

[5] I.e. using the daily logarithmic return data or the daily return data, to estimate the risk in relative terms.

3.3 Mathematical models to model missing data

If we do not have a price for a security today, the simplest assumption is that it is the same as it was yesterday. A slightly more sophisticated approach is to assume it has added a value to its price according to the drift (in the standard model). It is a form of linear extrapolation. Even more sophisticated is to simulate its value from the standard model, what we call Brownian extrapolation. A completely different question arises when we look back at a time series, and see that certain values are missing. Then we can say a missing price (if the prices immediately before and after are not missing) is simply the average of the prices captured on the days immediately before and after the missing value. This is linear interpolation; even this can be done more sophisticated, like with help of a Brownian Bridge if we assume the standard model (with fixed drift and volatility), or even more sophisticated with using time series modeling, which is beyond the scope of the present paper.

3.3.1 Linear models

Linear interpolation is principally easy: assume the prices Y_t are known days t and $t+k$, $k>1$, but unknown in between. Linear interpolation applied to the daily prices means that the estimated values are

$$\hat{Y}_{t+j} = Y_t + j\frac{Y_{t+k} - Y_t}{k}, \quad j = 1, \dots, k-1. \tag{3.11}$$

Linear extrapolation is another form of saying the last known value is used, but with adding the drift. Let us again assume that the index prices Y_t are known days t and $t+k$, $k>1$, but unknown in between. Formally, we define extrapolated values as

$$\hat{Y}_t = Y_t; \quad \hat{Y}_{t+j} = \hat{Y}_{t+j-1}e^{\mu}, \quad j = 1, \dots, k-1. \tag{3.12}$$

3.3.2 Standard Model and Brownian Bridge

Since the Brownian Bridge plays an important role for this paper, we describe here what it is, why it is used here and also the algorithm involved.

We assume that the index develops according to the standard model given by formula 2.3. In some cases, we shall write $W(t)$ instead of W_t to improve readability.

In this section, we will use the business day as unit. The drift is notoriously difficult to estimate, at least with both meaningful accuracy and statistical significance. The volatility changes: there are periods of higher and lower volatility on the market, the best we can do is to find a sensible volatility range and then investigate the effects with different choices of σ. We shall later show that as long as the volatility stays within sensible bounds it does not really matter for the purposes of this paper exactly how large it is. We shall therefore not investigate the problem of correctly estimating the present volatility further. We simply estimate both drift and volatility from the given time series.

Missing data and time lag requires essentially the same simulation tool in this respect.

If there is missing data somewhere in the middle of the time series, say k days are missing, then you can simulate the last missing day with help of the Brownian Bridge, noting that you are seeking $Y_{t+k-1}|Y_{t+k}$, then $Y_{t+k-2}|Y_{t+k-1}$, and so on until all missing values have been set.

Let τ be a time lag, $0 < \tau < 1$. To simulate from a time lag, we have to simulate from the conditional distribution $Y_{t+\tau}|Y_{t+1}$. The value of Y_t is known in both cases. So is the value at the end of the period. To further simplify the equations, we can assume that $t = 0$. Then both cases can be described as simulating from

$$Y_s|Y_u, 0 < s < u \tag{3.13}$$

or equivalently and more in detail,

$$Y_0 e^{\mu s+\sigma W(s)}|Y_0 e^{\mu u+\sigma W(u)}. \tag{3.14}$$

Here, we know the values of Y_0, s, u, and $\mu u + \sigma W(u)$, and suppose that the values of μ, σ and $W(u)$ are fixed, but not known to us. We have to simulate from $W(s)$. Of the values μ, σ and $W(u)$ we can choose two; the third can be calculated. Since we are going to go through this simulation for a given time series several times assuming the same μ and σ values, these two will be set and $W(u)$ be calculated.

To simulate from $W(s)$ we note that (see [26]) it has a normal distribution, and

$$\mathrm{E}[W(s)|W(u) = B] = \frac{s}{u}B \qquad \mathrm{Var}[W(s)|W(u) = B] = \frac{s}{u}(u - s). \tag{3.15}$$

A Brownian motion process within an interval, with its values fixed at both ends, is called a Brownian Bridge [12].

All that remains before we can begin the simulations using Brownian bridges is to choose the drift and volatility values. Drift is usually taken as a few percent; per day, as we have it in our time series, it changes the price of the asset (the index) with less than three hundredths of a percent. The volatility must lead to at least an order of magnitude bigger changes; there are very few days when the index does not change at least a few tenths of a percent (and quite often, more than one percent). This seems to imply that the drift is not critical for our investigations; we could even set it to zero, without affecting the outcome much.

The volatility is on the other hand is seemingly very important; even without calculating specific values; we can see that the bigger the volatility, the bigger the errors are likely to be.

Sometimes the missing data is at the beginning or at the end of the time series. Then no Brownian Bridge is needed, just the standard Wiener process.

The reasoning above gives the shortest, most transparent calculations. To see the effect of volatility changes without unnecessarily many simulations, we have to note how 3.15 can be modified if the volatility, i.e. the square root of variance changes. If we instead consider the Brownian motion process with $Var(W(t)) = \sigma^2 t$ on the unit interval and pin down both ends, the first to 0, we can replace 3.15 with 3.16 below:

$$\mathrm{E}[W(s)|W(u) = B] = \frac{s}{u}B \qquad \mathrm{Var}[W(s)|W(u) = B] = \sigma^2 \frac{s}{u}(u - s). \tag{3.16}$$

3.4 Time lags

Let τ be a time lag, $0 < \tau < 1$. The linearly interpolated tracking portfolio has a value which can be closely approximated as

$$\hat{Y}_{t+\tau} = e^{(\mu(t+\tau)+\sigma((1-\tau)W_t+\tau W_{t+1}))}. \tag{3.17}$$

Since we do not know the time lagged value of the index, we have to use a Brownian Bridge to set a value for it,

$$\tilde{Y}_{t+\tau} = e^{\left(\mu(t+\tau)+\sigma\left(W_t+\tau(W_{t+1}-W_t)+\sqrt{\tau(1-\tau)}u\right)\right)}, \tag{3.18}$$

where u has a standard normal distribution, in accordance with formulas 3.15 and 3.16. In other words,

$$\tilde{Y}_{t+\tau} = \hat{Y}_{t+\tau} e^{\left(\sigma\sqrt{\tau(1-\tau)}u\right)}. \tag{3.19}$$

Therefore

$$\log(\hat{Y}_{t+\tau}) = (\mu(t + \tau) + \sigma((1 - \tau)W_t + \tau W_{t+1})) \tag{3.20}$$

and

$$\begin{aligned} a_t &= \Delta\log(\hat{Y}_{t+\tau}) = \Delta\left(\mu(t+\tau) + \sigma((1-\tau)W_t + \tau W_{t+1})\right) \\ &= \mu + \sigma(\tau W_{t+2} - W_{t+1} + (1-\tau)W_t). \end{aligned} \tag{3.21}$$

And

$$\begin{aligned} b_t &= \Delta\log(\tilde{Y}_{t+\tau}) = \Delta\left(\mu(t+\tau) + \sigma\left((1-\tau)W_t + \tau W_{t+1} + \sqrt{\tau(1-\tau)}\right) u\right) \\ &= \mu + \sigma(\tau W_{t+2} - W_{t+1} + (1-\tau)W_t) + \sigma\sqrt{\tau(1-\tau)}(u-v), \end{aligned} \tag{3.22}$$

where u and v are independent standard normal variables.

We need the variances and the covariance of a_t and b_t.

We can assume without loss of generality that $t = 0$. The relative risk should be the same on the first day as on any other day. Then we can see that

$$\begin{aligned} Var[a_t] &= \sigma^2((1-\tau)^2 + \tau^2) \\ Var[b_t] &= \sigma^2((1-\tau)^2 + \tau^2 + 2\tau(1-\tau)). \end{aligned} \tag{3.23}$$

And finally, their covariance is

$$\begin{aligned} Cov(a_t, b_t) &= E[a_t b_t] - E[a_t]E[b_t] \\ &\approx \{\text{since } \mu \text{ is two order of magnitude smaller than } \sigma\} \\ &\approx E[\sigma^2\{\tau W_2 - W_1\}\{\tau W_2 - W_1 + \sqrt{\tau(1-\tau)}(u-v)\}] \\ &= \{\text{since } W_2 = W_1 + Z_1 \text{ independent}\} \\ &= \sigma^2 E[(\tau-1)^2 W_1^2 + \tau^2 Z_1^2] = \sigma^2((\tau-1)^2 + \tau^2) \end{aligned} \tag{3.24}$$

Combining 3.23 and 3.24, we can calculate $\hat{\sigma}^2 = \omega' \Sigma \omega$ for the variance-covariance method:

$$\begin{aligned} \hat{\sigma}^2 &= \omega' \Sigma \omega \\ &= [1 \quad -1] \begin{bmatrix} \sigma^2((\tau-1)^2 + \tau^2) & \sigma^2((\tau-1)^2 + \tau^2) \\ \sigma^2((\tau-1)^2 + \tau^2) & \sigma^2((1-\tau)^2 + \tau^2 + 2\tau(1-\tau)) \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix} \\ &= 2\tau(1-\tau)\sigma^2, \end{aligned} \tag{3.25}$$

which is quite close to the results achieved by simulations. For example, with half a day's time lag, this formula gives 0.0205 for the active portfolio, while simulation at average gave 0.0200.

How does interpolation with Brownian Bridges compare to this result? Obviously, b_t will not change, but a_t will since $\hat{Y}_{t+\tau}$ changes:

$$\hat{Y}_{t+\tau} = e^{\left(\mu(t+\tau) + \sigma\left((1-\tau)W_t + \tau W_{t+1} + \sqrt{\tau(1-\tau)}u'\right)\right)}, \tag{3.26}$$

where u' is a standard normal variable independent of the process W_t and from the process u. Consequently, $a_t = b_t$. Further calculations show that the covariance also remains the same, and so for the Brownian Bridge case, we have

$$\begin{aligned}\hat{\sigma}^2 &= \omega' \Sigma \omega \\ &= [1 \quad -1] \begin{bmatrix} \mathfrak{a}_{1,1} & \mathfrak{a}_{1,2} \\ \mathfrak{a}_{2,1} & \mathfrak{a}_{2,2} \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix} \\ &= 4\tau(1-\tau)\sigma^2, \end{aligned} \tag{3.27}$$

where

$$\begin{aligned}\mathfrak{a}_{1,1} = \mathfrak{a}_{2,2} &= \sigma^2((1-\tau)^2 + \tau^2) + 2\tau(1-\tau) \\ \mathfrak{a}_{1,2} = \mathfrak{a}_{2,1} &= \sigma^2((1-\tau)^2 + \tau^2). \end{aligned} \tag{3.28}$$

This shows analytically that there is nothing to gain from the Brownian Bridge, in fact, it deteriorates the result from linear interpolation with a factor of $\sqrt{2}$.

Similar calculations could be made for both extrapolation and Brownian extrapolation, but it is outside the scope. What remain is simulation, and a statistical analysis of the simulation results, to further underpin this conclusion. This will be done in section 5. There we shall perform simulations for all four possible algorithms, linear interpolation and extrapolation, and Brownian Bridge and Brownian extrapolation. The outcome of these simulations will also undergo hypothesis testing to show that indeed, the best we can do is linear interpolation.

3.5 The error term

Earlier in section 2.2, we have defined the error term as $e^{\varepsilon Z_t}$ with a sufficiently small ε and a standard normal variable Z_t. It is also straightforward to calculate the variance-covariance matrix for this case, using the well-known identities for variance and covariance, see e.g. [13]. Since for small values of x we have $e^x \approx 1 + x$, a small error in the data is adequately approximated with a small error in logarithmic return. In other words, we can use the approximation $\hat{Y}_t \approx Y_t(1 + \varepsilon Z_t)$.

Consequently, the value of the active portfolio is $V_t = \hat{Y}_t - Y_t \approx Y_t \varepsilon Z_t$. Since Y_t itself is a random variable, it is difficult to estimate the variance of V_t. The only thing that we can deduce from this approximation that the effect of small errors is proportional to ε.

If we had an additive error term, i.e. $\hat{Y}_t = Y_t + \varepsilon Z_t$, then we could make an explicit estimation of the effect of the error term. In this case, the variance-covariance matrix would have the value

$$\Sigma = \begin{bmatrix} \sigma^2 + \varepsilon^2 & \sigma^2 \\ \sigma^2 & \sigma^2 \end{bmatrix}. \tag{3.29}$$

Since the weight vector is $\omega_t = [1 - 1]'$ we can explicitly calculate the actual daily VaR value, using of course daily data for the drift as well:

$$\text{VaR}_\alpha = z_{1-\alpha}\sqrt{\omega_t'\Sigma\omega_t} - \omega_t\underline{\mu} = z_{1-\alpha}\varepsilon, \tag{3.30}$$

i.e. the variance to be used in the VaR calculation is simply ε^2 – if the error term is additive.

We could have arrived to the same result noting that the active portfolio should not have any drift, and its variance is approximately ε^2 so we can calculate the theoretical VaR value of the active portfolio for $\alpha = 0.99$ as

$$\text{VaR}_{0.99} \approx z_{0.01}\varepsilon. \tag{3.31}$$

3.6 Determining the volatility range

According to the standard model described earlier in equation 2.3, the price of an asset like the index of our choice can be modeled by the stochastic equation

$$Y_T = Y_t e^{\mu(T-t)+\sigma(W_T - W_t)}, \tag{3.32}$$

where μ and σ are deterministic constants, denoting the local rate of return and the volatility, respectively, and W_t is a standard Wiener process. So far in this paper we have taken μ and σ given, but they are not immediately observable at any time.

Their values have to be determined from the available data in a more or less roundabout way, much depending on what the model will be used for. Implied volatility is determined from the observable variables above, and the price of a derivative asset. Historic volatility is determined from a time series of historic asset prices. Here we use the volatility to simulate missing asset prices from a Brownian bridge. Historic volatility [4] seems to be a straightforward choice. It is calculated as follows.

At the $n+1$ equidistant points in time $t_0, ..., t_n$ we observe the index prices, and calculate the log-returns

$$\xi_i = \ln\left(\frac{Y_{t_i}}{Y_{t_{i-1}}}\right) \qquad i = 1, \ldots, n;\ \Delta t = t_i - t_{i-1}. \tag{3.33}$$

The sample variance S_ξ for the log-returns so calculated is

$$S_\xi^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\xi_i - \overline{\xi})^2, \tag{3.34}$$

where $\overline{\xi} = \frac{1}{n\Delta t}\sum_{i=1}^{n}\xi_n$ is the sample mean. Finally, the volatility is estimated as

$$\sigma^* = \frac{S_\xi}{\sqrt{\Delta t}}. \tag{3.35}$$

3.7 Hypothesis testing and P-value

Testing two methods against each other, such as linear interpolation and Brownian Bridge, we shall alway use samples with matched pairs. To compare the samples $\{a_1, \dots, a_n\}$ and $\{b_1, \dots, b_n\}$, we shall investigate the sample differences $\{x_1, \dots, x_n\} = \{a_1 - b_1, \dots, a_n - b_n\}$.

Throughout the paper, we are interested in comparing VaR values. All the simulations will be performed so that we can compare samples with matched pairs (of VaR values). Each time, we shall perform 100 simulations to obtain a suitably large sample size.

There are several ways of doing this, depending on how much we know of the distribution of the difference of the pairs in the sample. Tests for normality have shown that normality can not be taken for granted, so we use a simple sign test.

If there is no difference between the samples, the median of their difference should be 0. This means that for the samples $\{a_1, \dots, a_n\}$ and $\{b_1, \dots, b_n\}$

$$P(a_i - b_i > 0) = \tfrac{1}{2} \quad \text{where } i = 1, \dots, n.$$

Now assume that we have k pairs, $k < \frac{n}{2}$, where $a_i > bi$. What is the probability P_k that we have a result that is at least this extreme? This is the question a two-sided simple sign test [5] answers. Observe that this test is not specific about the distribution of differences; it only requires that the mean should exist.

The result would be equally extreme if we had $n-k$ pairs or more where $a_i > b_i$, or k pairs or less where $a_i > b_i$.

The probability that $x_i > y_i$ for exactly j pairs is

$$p_j = \binom{n}{j}\left(\frac{1}{2}\right)^j\left(\frac{1}{2}\right)^{n-j} = \binom{n}{j}\left(\frac{1}{2}\right)^n \tag{3.36}$$

so

$$P_k = \sum_{j=0}^{k} \binom{n}{j} \left(\frac{1}{2}\right)^n + \sum_{j=n-k}^{n} \binom{n}{j} \left(\frac{1}{2}\right)^n. \qquad (3.37)$$

Let the null hypothesis be that there is no difference, i.e. if we denote the mean with m,

H_0: $m = 0$

and the alternative hypothesis

H_1: $m \neq 0$.

When we test for difference between two matched pair samples, P_k will be the observed level of significance or P-value. This means that for any given α we can reject H_0 on the significance level α if $\alpha > P_k$.

The program we use to calculate the observed level of significance, MATLAB, sometimes only shows this value to four decimal places. This means that the value $P_k = 0$ means only that $P_k < 0.00005$.

This was the two-sided test. If on the other hand we had

H_0: $m \leq 0$

and

H_1: $m > 0$.

then

$$P_k = \sum_{j=0}^{k} \binom{n}{j} \left(\frac{1}{2}\right)^n \qquad (3.38)$$

i.e. the observed level of significance would be exactly half as much as in the case of a two-sided test.

4 Data

The same data set and the same algorithms can lead to different results depending on exactly how the data is used.

4.1 Market data set

The first and obvious use of the given data sets (OMXS30 data for one year, tracker portfolio value for one year, 2010) is to calculate the drift and the volatility of this index and to calculate the VaR for the active portfolio using the variance-covariance method. The drift and volatility will be used in subsequent calculations. All three sources of error need their own analysis and consideration. As before, the daily average return or daily drift will be denoted μ and the daily variance or daily volatility with σ.

4.2 Missing data

If there is just one day of missing data, and we are interested in the VaR on that day, the daily drift and volatility provides the distribution for the logarithmic return even without simulation.

First consider linear interpolation from adjacent values. Then we tacitly supposed that the tracker portfolio coincided with the index the day before. Using first order approximation to e^x we can write

$$\hat{Y}_t \approx Y_0 e^{\mu t + \sigma\left(\frac{1}{2}W_{t+1} + \frac{1}{2}W_{t-1}\right)} \tag{4.1}$$

and estimate the actual loss from having to estimate the missing data as

$$\begin{aligned} L \approx Y_t - \hat{Y}_t &= Y_0 e^{\mu t}\left(e^{\sigma W_t} - e^{\sigma\left(\frac{1}{2}W_{t+1} + \frac{1}{2}W_{t-1}\right)}\right) \\ &= Y_0 e^{\mu t + \sigma W_t}\left(1 - e^{\sigma\left(\frac{1}{2}W_{t+1} - W_t + \frac{1}{2}W_{t-1}\right)}\right) \\ &= Y_t\left(\sigma\left(\frac{1}{2}(W_{t+1} - W_t) - \frac{1}{2}(W_t - W_{t-1})\right)\right). \end{aligned} \tag{4.2}$$

In formula 4.2 we have the daily changes in paranthesis, and those are independent of each other. Their variances add upp. If X is a random variable and c is a constant, $Var[cX] = c^2 Var[X]$. Therefore the variance to be used in the relative VaR calculation is

$$\begin{aligned} &Var\left[\sigma\left(\frac{1}{2}(W_{t+1} - W_t) - \frac{1}{2}(W_t - W_{t-1})\right)\right] \\ &= \sigma^2\left(\frac{1}{4} + \frac{1}{4}\right) = \frac{1}{2}\sigma^2. \end{aligned} \tag{4.3}$$

If we simply put the last day's index price into today's tracker portfolio price, then we have

$$V_t = \hat{Y}_t - Y_t = Y_{t-1} - Y_t. \tag{4.4}$$

Noting that the value of the active portfolio was in this case 0 on the previous day, we can conclude that it has exactly the same VaR as the index itself, i.e. σ^2 – which is obvious even without calculations.

Now let us consider a more sophisticated interpolation, that with a Brownian Bridge. Is it more efficient? The value of the active portfolio on a day when there is no data, but when there is data on the adjacent days, is

$$V_t = \hat{Y}_t - Y_t = e^{\mu t + \sigma \hat{W}_t} - e^{\mu t + \sigma W_t} \approx \sigma(\hat{W}_t - W_t). \tag{4.5}$$

The quantity $\hat{W}_t$ is calculated by simulating from a Brownian Bridge. To get an explicit expression, note that $\hat{W}_t = W_{t-1} + \hat{B}_1$ and $W_t = W_{t-1} +$

$(W_t - W_{t-1}) = W_{t-1} + B_1$ where B is a standard Wiener process. $\hat{B}_1$ is then a (standard) Brownian Bridge, with $s = 1$ and $u = 2$ using the notation in formula 3.15. This means it has an expected value of $\frac{1}{2}(W_{t+1} - W_{t-1})$ and a variance of $\frac{1}{2}$. This means we can express $\hat{W}_t - W_t$ the following way:

$$\begin{aligned} \hat{W}_t - W_t &= \left(W_{t-1} + \frac{1}{2}(W_{t+1} - W_{t-1}) + \frac{1}{\sqrt{2}}\tilde{B}_1 \right) - W_t \\ &= \frac{1}{2}(W_{t+1} - W_t) - \frac{1}{2}(W_t - W_{t-1}) + \frac{1}{\sqrt{2}}\tilde{B}_1, \end{aligned} \tag{4.6}$$

which in turn means that the compound variance is exactly σ^2.

Extrapolation with the standard model will result in an even larger variance, $2\sigma^2$.

This answers the question what effect one day of missing data has just on the day in question. A completely different question is what effect one day of missing data has on the expected daily VaR. To answer this question analytically is not so easy, since the resulting distribution is far from being normal. We have to resort to simulation, and use the variance-covariance method with parameters estimated from the values the simulation provides.

While it is true that even this calculation builds on a premise which is not true, namely that the active portfolio as constructed has a joint normal distribution, it can be argued that this is a good approximation. We have used the index, generated missing data days and replaced the index value with generated values.

Even if we do not have an analytical result for the expected daily VaR in case of missing data, the calculations in this section suggest that the smallest VaR comes from linear interpolation.

4.3 The volatility range

The historic values of OMXS30 are available from the NASDAQ OMX website for the last 20 years. For a bigger sample, we can calculate the volatility of the AFGX for which there is readily available data (at least monthly) from 1949 to 2004 (at the website of the Swedish periodic Affärsvärlden); that gives another forty-one volatility figures, if we assume that these indices have about the same volatility range, and choose only one value for the years when both are available.

For the years we have both OMXS30 and AFGX, the daily volatility ranges of these two indices are:

$$\{\sigma_{OMXS30}|0.006 \leq \sigma_{OMXS30} \leq 0.021\}$$

and

$$\{\sigma_{AFGX}|0.006 \leq \sigma_{AFGX} \leq 0.021\}$$

respectively; the same to the third decimal point. For the longer time frame, the AFGX volatility range is

$$\{\sigma_{AFGX}|0.003 \leq \sigma_{AFGX} \leq 0.024\}.$$

Considering this, we can say that the sensible volatility range to use is

$$\{\sigma|0.003 \leq \sigma \leq 0.03\};$$

the histogram in Figure 1 suggests (although not proves in any way) that volatility comes from a log-normal distribution, in accordance with the findings in [20].

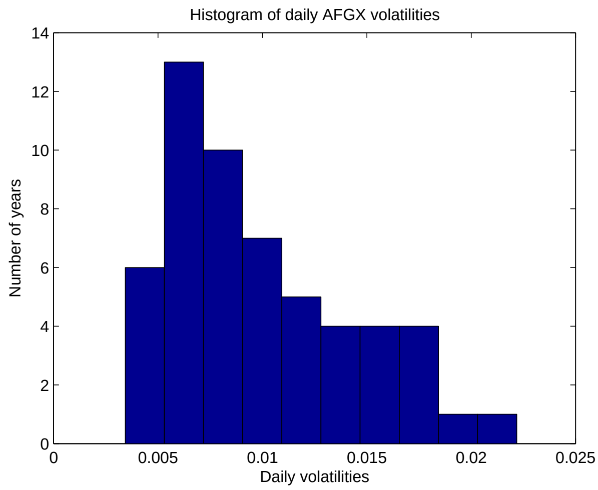

Figure 1: Histogram depicting the daily volatilities of the AFGX index, calculated from monthly data, for 1949-2004.

5 Simulations

5.1 First round: missing data, linearly interpolated

In the first round of simulations we have varied the parameters for the Markov chain, p_{00} went from 0.5 to 1, while p_{11} varied from 0.9 to 1, all in 0.01 increments. The confidence level remained the same, 0.99, throughout these simulations.

The result is 561 (i.e. 51 times 11) $\mathrm{VaR}_{0.99}$ values, corresponding to the possible combinations of p_{00} and p_{11} values.

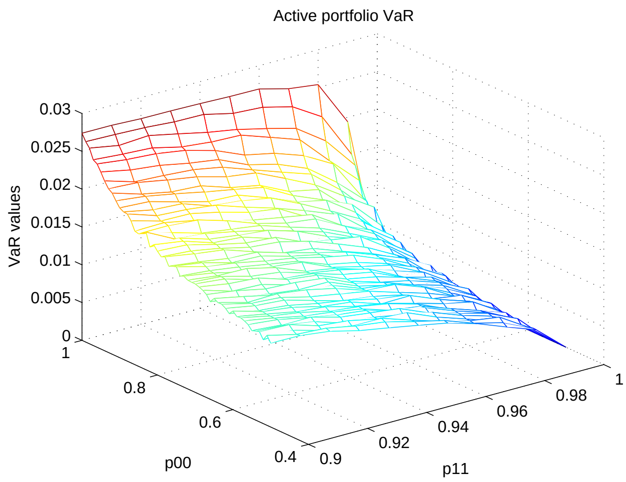

Figure 2: Hundred simulated VaR values, linear interpolation.

Figure 2 tells the story about how the transition probabilities work. It shows that the relative VaR_αvalue for $p_{11} = 0.9$ and $p_{00} = 1$ is circa 0.0275, and that the VaR_α value for $p_{11} = 0.9$ and $p_{00} = 0.5$ is around 0.01. Of course if the probability that we generate another missing data day after a missing data day is high, VaR will also be high. Even if the probability of getting missing data after a day with data is high, i.e. the probability of getting another day with data is low, will the VaR be high. If a missing data day means there is no more data in the rest of the period, then VaR will be about the same order of magnitude as e.g. the tracker portfolio. It is interesting to note that the effect of p_{00} is more marked than the effect of p_{11} as Figure 2 shows.

Linear interpolation is ”nicer” than the Brownian Bridge; it always ends up between the neighboring values. On the other hand, the stock market does not behave that nicely. It is then interesting to see what generating the missing values by means of Brownian Bridges changes. Earlier, calculations in sections 3.4 and 4.2 have shown that at least for time lags and for one day of missing data, linear interpolation resulted in a lower VaR value than interpolation with Brownian Bridges. The design of the Brownian Bridge is principally adding an independent random variable to the value calculated by linear interpolation. It results therefore in a larger variance (than linear interpolation itself), and (if variance is the dominating factor in the calculation of the VaR) a larger VaR. In the next section, we shall show by means of simulation that this indeed is the case even for missing days.

5.2 Second round: missing data, using Brownian Bridge

To make the results easier to compare, we have used the same missing values as in the previous simulations, but now Brownian Bridges bridged the gaps instead of linear interpolation. The general shape of the surface describing the effects of varying transition probabilities in Figure 3 is about the same as in Figure 2, but the magnitude is different. For example, the value for $p_{11} = 0.9$ and $p_{00} = 1$ is circa 0.0395. It is very close to the value which we obtain if we multiply the corresponding value from Figure 2 with $\sqrt{2}$ (which would yield 0.0389). We have seen when analysing time lag effects that there the multiplier between linear interpolation and Brownian Bridge was indeed $\sqrt{2}$. Is there a similar connection in this case also? We have

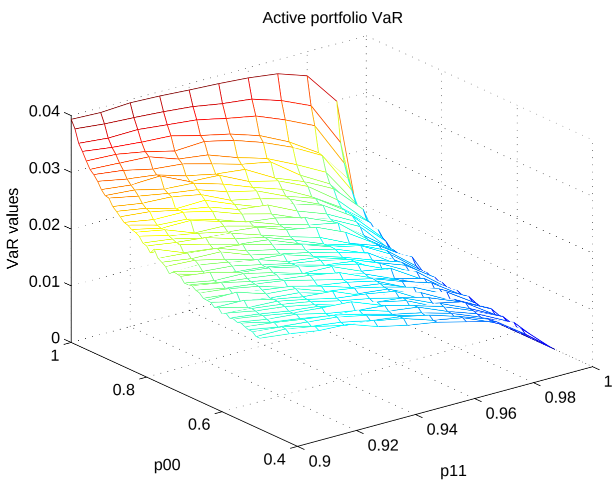

Figure 3: Hundred simulations with Brownian Bridges, with historical volatility from the sample (circa 1.2 %).

made two sets of statistical tests. In total, there were 224400 simulations made[6]: for each pair of p_{11} and p_{00} values, we made 100 Markov chains, and for each Markov chain, we have simulated from all four cases.

First set of statistical tests We have first wanted to see if linear interpolation is different from Brownian Bridge interpolation. For this purpose, we have made a two-sided test for all 550 interesting pairs (of course, when $p_{11} = 1$, there is not much to compare). Each of this pairs utilised hundred simulated Markov chains. For each Markov chain, we

[6]Each point on the surface, in e.g. Figure 2 or Figure 3, represents a mean value of 100 simulated VaR values and there are 561 such points in each simulation, i.e. in linear interpolation, Brownian Bridge, linear extrapolation and Brownian extrapolation.

have calculated the interpolated tracker portfolio and the Brownian Bridge tracker portfolio. The null hypothesis was that (for the par of values within one test run, i.e. the same Markov chain) there is no difference, i.e. that the difference is 0.

H_0: There is no difference between linear interpolation and Brownian Bridge.

H_1: There is a difference between linear interpolation and Brownian Bridge.

First, we have tested if the distribution of the differences in VaR values were normally distributed. We have used the Jarque-Bera [16] test for this purpose. We will not in depth describe the Jarque-Bera test in this paper, but rather briefly present an overview.

This test is a two-sided goodness-of-fit test suitable when a fully-specified null distribution is unknown and its parameters must be estimated. The test statistic is

$$JB = \frac{n}{6}\left(s^2 + \frac{(k-3)^2}{4}\right), \tag{5.1}$$

where n is the sample size, s is the sample skewness, and k is the sample kurtosis. For large sample sizes, the test statistic has a χ^2-distribution with two degrees of freedom.

In testing for normality, the null hypothesis that the sample comes from a normal distribution with unknown mean and variance, against the alternative that it does not come from a normal distribution. In more than 30% of cases, the result was that one could reject the normality hypothesis on a 99% confidence level.

Since the differences between the VaR values were not normally distributed in about a third of the cases, we have used a two-sided simple sign test to see if they are significantly different (see e.g. [5]).

All these differences were significant on at least 99.9% confidence level. In fact, even the least significant result showed a significance level under 10^{-6}. Having established that there indeed is a difference, we have made a one-sided test to see if the null hypothesis

H_0: $\mathrm{VaR}_{\mathrm{BB}} \leq \mathrm{VaR}_{\mathrm{lin\ int}}$

could be rejected on the 99.9% confidence level.

The answer was affirmative in all the 550 interesting cases (i.e. when p_{11} was different from 1).

Second set of statistical tests The next question was if we can reject the hypothesis that indeed the Brownian Bridge entails a $\sqrt{2}$ times worse result, hence

$$H_0\colon\ \sqrt{2}x_{\text{lin int}} - x_{\text{BB}} = 0$$

$$H_1\colon\ \sqrt{2}x_{\text{lin int}} - x_{\text{BB}} \neq 0$$

There we could not reject the null hypothesis on at least 99.9% confidence level in the majority of cases. The connection between these two methods is not as simple as it is for the time lag case. A detailed investigation of this relation is outside the scope of this paper.

To summarize, statistical tests proved that linear interpolation provides lower VaR values than using the Brownian Bridge method in the missing data case.

What happens if the volatility is higher? Of course it only matters for the Brownian Bridge, the interpolation does not change. To see the overall effect we have also included simulations (with the same Markov chains) using ten times the variance, hence $\sqrt{10}$ times the volatility, in the given index series. The result is depicted in Figure 4. It is clear from comparing Figures

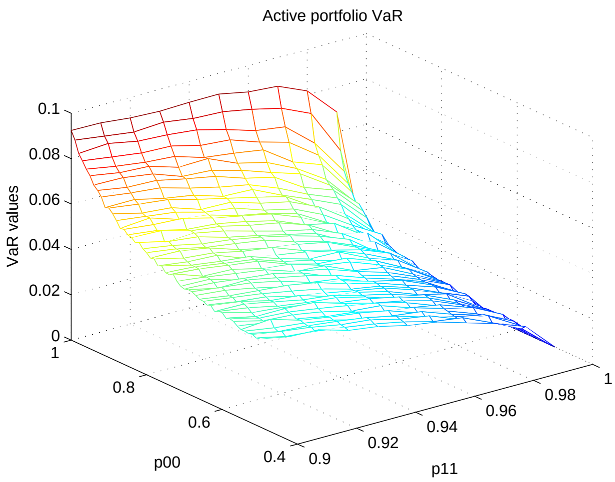

Figure 4: Missing data with Brownian Bridge using $\sqrt{10}$ times the historical volatility.

3 and 4 that higher volatility results in higher VaR values as expected, but is the effect proportional to σ? Noting that $0.0395\sqrt{10} = 0.1249$ and not 0.92 (as the upper right corner of Figure 4 would suggest) shows that proportionality does not seem to hold, at least not if we change the volatility

but use the same index series. Still, the question of how the volatility used in the calculations affects the results is an interesting one.

To see at least some of the effects, we fix the transition probabilities[7] and let σ vary. In Table 1, the first column contains the volatility, the second column contains the linearly interpolated VaR values (at average), while the third column shows Brownian Bridge interpolated VaR values (at average). In all cases, the difference between the last two columns was significant at the 99.9% confidence level. The volatility does not come in play with linear

Volatility	**Linear Interpolation**	**Brownian Bridge**
0.0030	0.0087	0.0089
0.0060	0.0090	0.0100
0.0090	0.0090	0.0114
0.0120	0.0088	0.0126
0.0150	0.0087	0.0138
0.0180	0.0088	0.0161
0.0210	0.0085	0.0172
0.0240	0.0089	0.0196
0.0270	0.0093	0.0217
0.0300	0.0088	0.0237

Table 1: The effect of volatility.

interpolation; the differences there come from the different outcomes from simulating the Markov chain.

It is difficult to see from a table if the effect is linear. We have therefore included a graph, Figure 5, where we have plotted VaR values, resulted from the simulation using Brownian Bridge, as a function of volatility. As we can see, the effect is not strictly linear, but quite close to being linear[8].

There is one more area to investigate. What if there is only a sporadic day of missing data due to e.g. clerical error, and those days come independently of each other? These conditions can be realised if $p_{11} = p_{01}$ (and consequently $p_{10} = p_{00}$) and p_{11} is relatively high. To make it a little more general, we let both p_{11} and p_{01} vary between 0.98 and 1, in 0.001 increments. Up until $p_{11} = 0.994$ the linear interpolation resulted in significantly lower VaR values than Brownian Bridges, on a 99.9% confidence level, independently of what p_{01} was chosen. Then for $p_{11} = 0.994$ and $p_{11} = 0.995$ there were two instances each (i.e. 2 of 20) where the confidence level dropped

[7] We set $p_{11} = 0.9$ and $p_{00} = 0.1$.

[8] Investigating linearity is not in the scope of this paper.

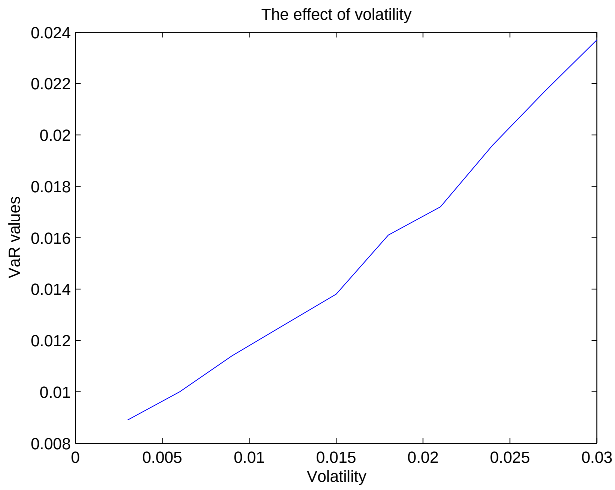

Figure 5: The effect of varying the volatility.

under 99.9%. For $p_{11} = 0.996$ and for higher p_{11} values, the results were inconclusive.

We have used one year long time series, 252 days. In terms of days – $p_{11} = 0.994$ means one and a half missing days per year at average; $p_{11} = 0.995$ means a little over one day. Higher p_{11} values mean not even one day per year missing at average. It is natural that the results are then thrown by single days.

5.3 Third round: missing data, linearly extrapolated

Figure 6 shows average VaR values for missing data with extrapolation. Compare with Figures 2 and 3, containing missing data simulations with linear interpolation and Brownian Bridge. It is not obvious form the figures what the relations between these three methods are. We have performed the same kind of statistical analysis (two sided simple sign test) as we used to compare linear interpolation and Brownian Bridge. The VaR values calculated from linear extrapolation were in all 550 cases significantly higher than the values calculated from linear interpolation, as expected.

More surprisingly, the VaR values from linear extrapolation were lower than those obtained using the Brownian Bridges, in a majority of cases.

5.4 Fourth round: missing data, Brownian motion

We have not found a straightforward analytical description which would describe our four ways of providing missing data. We have expected to find

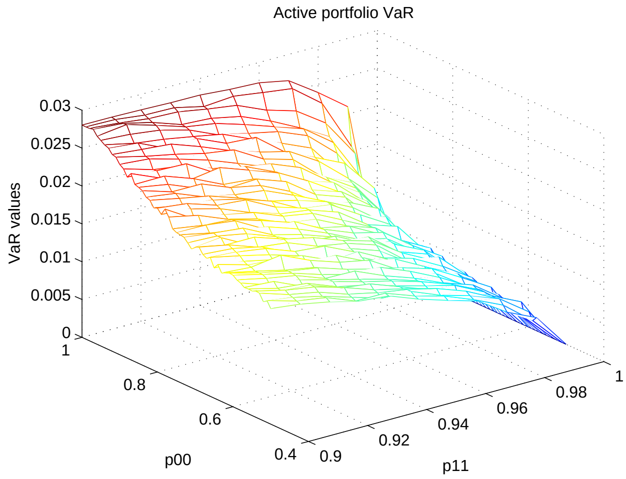

Figure 6: Hundred simulated VaR values, linear extrapolation.

the same kind of difference between linear extrapolation and Brownian extrapolations as we have found between linear interpolation and Brownian Bridge. Figure 7 displays the results from Brownian extrapolation.

Indeed, the same kind of statistical analysis, as in sections 5.2 and 5.3, showed that linear extrapolation resulted in lower VaR values than Brownian extrapolation, and the differences were significant at least on the 99.9% confidence level in each of the 550 simulated matched pair samples.

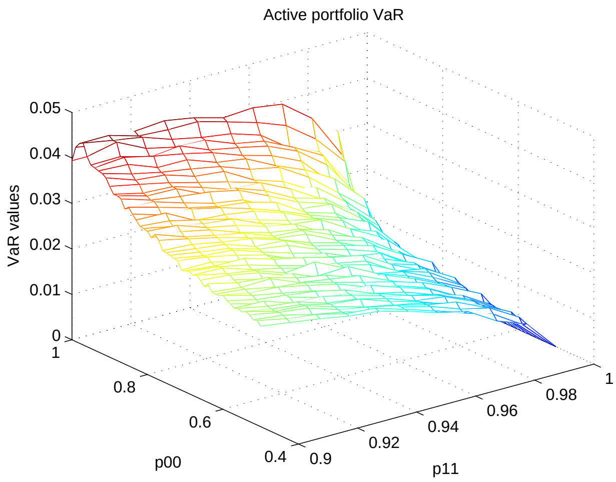

Figure 7: Missing data, Brownian extrapolation.

5.5 Fifth round: time lag

We have previously in section 3.4 calculated the effects of time lags using linear interpolation and Brownian Bridge. We have also performed simulations to show what the results really look like. These results are displayed in Tables 3-6, but first take a look at Table 2, where we present estimated volatility and VaR for the OMXS30 index years 2008-2010.

Year	σ	**VaR**$_\alpha$
2008	0.02505	0.05828
2009	0.01825	0.04246
2010	0.01216	0.02828

Table 2: OMXS30 volatility and VaR years 2008-2010.

Linear Interpolation			
Year	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$
2008	0.0359	0.0419	0.0357
2009	0.0262	0.0300	0.0259
2010	0.0174	0.0202	0.0173

Table 3: VaR calculated for the active portfolio varying τ values using Linear Interpolation years 2008-2010.

Linear Extrapolation			
Year	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$
2008	0.0415	0.0593	0.0725
2009	0.0303	0.0428	0.0530
2010	0.0202	0.0287	0.0350

Table 4: VaR calculated for the active portfolio varying τ values using Linear Extrapolation years 2008-2010.

Brownian Bridge			
Year	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$
2008	0.0506	0.0585	0.0511
2009	0.0368	0.0423	0.0367
2010	0.0246	0.0285	0.0245

Table 5: VaR calculated for the active portfolio varying τ values using Brownian Bridge years 2008-2010.

We have simulated from the data for three consequtive years, 2008-2010. The goal was in part to compare even with simulation which methods are best in estimating the time lagged value, in part to see what the effects are. Not surprisingly, using linear interpolation gives the best result. That linear

Brownian Extrapolation			
Year	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$
2008	0.0580	0.0837	0.1020
2009	0.0424	0.0602	0.0738
2010	0.0285	0.0399	0.0490

Table 6: VaR calculated for the active portfolio varying τ values using Brownian Extrapolation years 2008-2010.

interpolation gave the best result is also confirmed statistically. The differences between linear interpolation and any of the other three were significant on the 99.9% confidence level. More precisely, both the null hypothesis that the other methods are as good as linear interpolation, and the null hypothesis that they are at least as good as linear interpolation could be rejected at the 99.9% confidence level.

The simulation results are quite close to the predictions, when such could be made from our previous calculations, i.e. the results from applying equations 3.25 and 3.27 to e.g. equation 3.10 (either with the appropriate μ values, or taking $\mu = 0$).

Only considering 2010 data, the difference between linear extrapolation and Brownian Bridge for $\tau = 0.5$ seemed to be significant. However, in the other years, it was not. Table 7 summarizes the results of significance testing. Note that types of test in Table 7 are ranked in such a way that tests

Type of test	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$
Brownian Bridge vs. Linear Inter	***	***	***
Linear Extra vs. Linear Inter	***	***	***
Brownian Extra vs. Linear Inter	***	***	***
Linear Extra vs. Brownian Bridge	(***)	—	***
Brownian Extra vs. Brownian Bridge	***	***	***
Brownian Extra vs. Linear Extra	***	***	***

Table 7: The significances of differences between methods in the time lag case.*** means significance on the 0.001 level. — means no significance. (***) means significant but in opposite direction.

with VaR being significantly higher are ranked first and then in a decreasing order[9].

[9] I.e. for Brownian Bridge compared to Linear interpolation first, followed by Linear extra compared to Linear interpolation etc.

5.6 Sixth round: normally distributed small errors

Normally distributed small errors of the portfolio values correspond to a good precision to normally distributed small errors of the logarithmic returns. Therefore, we just used the logarithmic returns and attached an additive error term.

Making the simulation, we get the results illustrated in Figure 8. For

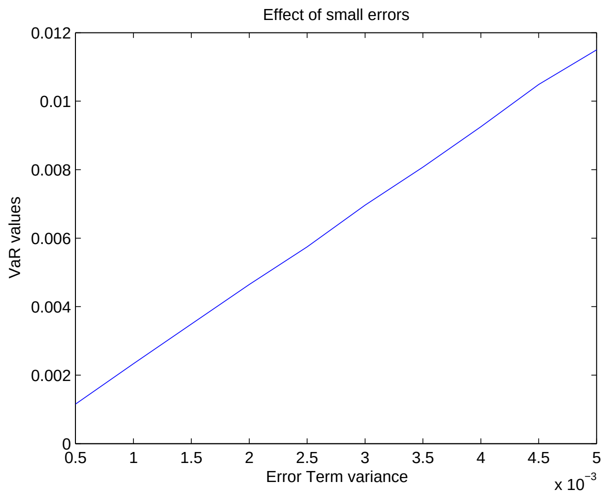

Figure 8: The effect of small errors.

$\varepsilon = 0.005$ the daily relative VaR from simulations was 0.0117, actually very close to the theoretical value of an additive error term to the normed index (i.e. index which begins with the value of 1), which is 0.0116.

5.7 Seventh round: other historic data

In section 5.2 we have briefly investigated what happens if we change the volatility used in the simulations. The historic volatility of the data remained the same. Now we shall also investigate what happens if the historical volatility changes. As linear interpolation provided significantly lower VaR values than any other approach, we shall do it using linear interpolation only.

We have downloaded data about the OMXS30 index for 20 years[10], 1991-2010, and calculated the volatilities (see Table 8). As we can see, the highest (daily) volatility occurred in 2008, when it was 0.025. The year with the lowest daily volatility was 2005 with 0.008. Historically, it was a very turbulent twenty years period.

[10] Data was downloaded from www.nasdaqomxnordic.com.

Year	**Volatility**
1991	0.013
1992	0.016
1993	0.011
1994	0.011
1995	0.009
1996	0.009
1997	0.014
1998	0.020
1999	0.013
2000	0.020
2001	0.022
2002	0.022
2003	0.014
2004	0.010
2005	0.008
2006	0.012
2007	0.013
2008	0.025
2009	0.018
2010	0.012

Table 8: Historical OMXS30 daily volatilities, years 1991-2010.

To see the effect of historical volatilities (i.e. the prevailing volatilities, as opposed to volatilities used in calculation, which are always a guesswork), we choose three years, 2008, 2009, and 2010. Their volatilities were 0.025, 0.018 and 0.012 respectively, pretty much an even series. We have seen earlier that

Year	Missing data $p_{11} = 0.9$ $p_{00} = 0.5$	Missing data $p_{11} = 0.99$ $p_{00} = 0.5$	Missing data $p_{11} = 0.9945$ $p_{00} = 0.0055$	Half day's time lag	Small errors $\varepsilon = 0.001$
2008	0.0243	0.0067	0.0031	0.0419	0.0023
2009	0.0174	0.005	0.0029	0.0300	0.0023
2010	0.0115	0.0034	0.0015	0.0202	0.0023

Table 9: Effects of missing data, time lag and small errors when calculating VaR for the active portfolio.

the effects of time lags (calculated or simulated by linear interpolation) are roughly proportional to historic volatility. It seems logical that the effect of missing data and half a day's time lag (for a fixed set of other parameters) should also vary proportional to the volatility. Table 9 indeed indicates that such is the case, except for when we have very few missing days.

6 Conclusions

Decision makers need short, concise and easy to understand summaries. It is because they have to collate information from many sources and different specialists. Here we attempt to summarize the effects of sources of errors, and provide some simple rules. In the table which summarizes all the calculations in a comprehensive yet easy to understand way, we have included three cases of missing data, each with 100 simulations and taking the average; time lag with linear interpolation; and the effect of small errors. Analyzing Table 9 we can see that the relation between some of the columns is quite stable (not to the last decimal, but in most cases, with around a tenth relative precision). We can derive a small number of simple, but potentially useful rules from here.

Rule 1 The effect of time lags and missing data where the probability of a missing day is not too small is proportional to the volatility. If there are less than two missing days per year at average, then there is no such proportionality.

Rule 2 The effect of small errors on the other hand is proportional to the standard error of the white noise (or error term).

Rule 3 The effect of time lag τ is a multiplicative factor $\sqrt{2\tau(1-\tau)}$ on the VaR of the index, if we can use linear interpolation.

Rule 4 Introducing a new source of uncertainty increases VaR. For example, if we replace linear interpolation with Brownian Bridge interpolation, the calculated VaR increases.

7 Further Research

Although the Basel and UTICS frameworks emphasize VaR, other risk measures such as the Expected Shortfall can be more adequate in certain situations. Regulatory regimes come and go; even now there is a critic of the Basel II framework, for its potentially destabilizing effects on the very markets it aims to stabilize [14]. It may therefore be prudent to investigate other risk measures, perhaps based on other, more fat-tailed distributions.

Another possibility is to systematically compare empirical VaR (and other measures) to their parametric counterparts.

A third question is how much the assumptions of normality and independence really matter. Clearly, the active portfolio does not possess a two-dimensional normal distribution in the missing data case, although one can argue that normality holds for the other cases (provided logarithmic returns are normally distributed). With the introduction of Markov chains we have introduced a measure of temporal dependence. A more thorough analysis of what happens would perhaps involve GARCH models [6]. Before making such an analysis, however, we should investigate if just randomly distributing a number of missing data days would not give essentially the same result as the Markov chain based approach. The less assumptions we make, the more robust are the results.

We have seen in this paper that linear interpolation leads to the lowest VaR values for the active portfolio. We have derived analytical expressions for the effects of time lags and one day of missing data using the historical volatility. No such expression was derived for missing data cases in general. To see the effect of prevailing volatility on missing data VaR in general, one could make the same calculations for a number of years (i.e. for time series with different historical volatilities). To make the results comparable, one should use the same Markov chains for all those years.

When we in section 5.2 plot VaR values simulated using Brownian Bridge, as a function of volatility – the effect is quite close to being linear. However to investigate this further, a test on linearity can be performed.

References

[1] Al Janabi M. (2008): *On the appropriate function of trading risk management units: Primary roles and rational use of internal models.* Journal Of Banking Regulation [serial online], 10(1), pp. 68-87

[2] Beasley J.E., Meade N., Chang T.J. (2003): *An evolutionary heuristic for the index tracking problem.* European Journal of Operational Research 148, pp. 621-643

[3] Bingham N.H., Kiesel H. (1998): *Risk-neutral valuation : pricing and hedging of financial derivatives.* Springer Finance, Springer, London

[4] Björk T. (2004): *Arbitrage Theory in Continuous Time.* Oxford University Press, New York, Second Edition

[5] Blom G., Holmquist B. (1998): *Statistikteori med tillämpningar.* Studentlitteratur, Lund, Tredje utgåva

[6] Brockwell P. J., Davis R. A. (2002): *Introduction to Time Series and Forecasting.* Springer Texts in Statistics, Springer Verlag, New York, Second edition

[7] Colwell D., El-Hassan N., Kwon Oh Kang (2007): *Hedging diffusion processes by local risk minimization with applications to index tracking.* Journal of Economic Dynamics and Control, Vol. 31, Issue 7, pp. 2135-2151

[8] Crowder G. B., Kazemi H., Schneeweis T. (2010): *Asset Class and Strategy Investment Tracking Based Approaches.* Journal Of Alternative Investments, 13(3), pp. 81-101

[9] Embrechts P., Hofert M. (2011): *Practices and issues in operational risk modeling under Basel II.* Lithuanian Mathematical Journal, Vol. 51, No. 2, pp. 180-193

[10] Escanciano J. C., Olmo J. (2010): *Parametric Value-at-Risk With Estimation Risk.* Journal of Business & Economic Statistics, Vol. 28, No. 1

[11] García F., Guijarro F., Moya I. (2011): *The curvature of the tracking frontier: A new criterion for the partial index tracking problem.* Mathematical & Computer Modelling, 54(7/8) pp. 1781-1784

[12] Glasserman P. (2004): *Monte Carlo Methods in Financial Engineering.* Springer-Verlag, New York

[13] Gut A. (1995): *An intermediate Course in Probability.* Springer Verlag, New York

[14] Hermsen O.O. (2010): *Does Basel II destabilize financial markets? An agent-based financial market perspective.* European Physical Journal B – Condensed Matter, 73(1) (2010) 29-40

[15] Höglund T. (2008): *Mathematical Asset Management.* John Wiley & Sons, Hoboken, New Jersey

[16] Jarque C.M., Bera A.K. (1987): *A test for normality of observations and regression residuals.* International Statistical Review, Vol. 55, No. 2

[17] Jiménez-Martín J., McAleer M., Pérez-Amaral T. (2009): *The ten commandments for managing Value at Risk under the Basel II accord.* Journal Of Economic Surveys, 23(5), pp. 850-855

[18] Jorion P. (2003): *Financial Risk Manager Handbook*, John Wiley & Sons, Inc., Hoboken, New Jersey, Second edition

[19] Lolman J., Onak M., Jorion P., Taleb N., Derman E., Putnam B., Sandor R., Jonas S. et al (1998): *Roundtable: The Limits of VaR*, Derivatives Strategy April 1998

[20] Liu Y., Gopikrishnan P., Cizeau P., Meyer M., Peng Chung-Kang, Stanley H. Eugene (1999): *Statistical properties of the volatility of price fluctuations.* Physical Review, Vol. 60, No. 2

[21] Maspero D., Saita F. (2005): *Risk management for asset managers: A test of relative VaR.* Journal Of Asset Management, 5(5), pp. 338-350

[22] McGill Ross & Sheppey Terence (2005): *The new global regulatory landscape: Impact on finance and investment.* Basingstoke, Palgrave Macmillan

[23] McNeil A.J., Frey R., Embrechts P. (2005): *Quantitative Risk Management, Princeton Series in Finance.* Princeton University Press, Princeton, New Jersey

[24] Miao J., Dunis C. (2005): *Volatility Filters for Dynamic Portfolio Optimization.* Applied Financial Economics Letters, 1, pp. 111-119

[25] Raubenheimer H. H. (2003): *Serial correlation and TEV bias in index funds.* South African Journal Of Business Management, 34(2), p. 45

[26] Ross S.M. (2003): *Introduction to probability models.* Elsevier, Academic Press Burlington, Eight edition

[27] Salmon J. (2009): *Robots wage war on funds*, Daily Mail, London

[28] Simons K. (2000): *The Use of Value at Risk by Institutional Investors.* New England Economic Review [serial online], Available from: Business Source Premier, Ipswich

[29] Tamhane A.C., Dunlop D.D. (2000): *Statistics and Data Analysis: from elementary to intermediate.* Prentice Hall, Inc., Upper Sadle River

[30] Walker G. (2011): *Basel III market and regulatory compromise.* Journal of Banking Regulation

[31] Basel II:
http://www.imf.org/external/np/pp/eng/2005/072205.htm#s1

[32] NASDAQOMX exchange:
https://indexes.nasdaqomx.com/IndexMethodologies.aspx

[33] Riksbanken:
http://www.riksbank.com/templates/SectionStart.aspx?id=10604