Time Series Trading Models and Assessment

1 Introduction

This report will focus on the analysis and evaluation of three trading strategies: trend-following, mean-reverting and autoregressive. They will be applied to data generated with a linear time series function (defined below). Parameters for each strategy will be optimised in the training set of the data and performance will then be computed on the test set. The strategies will be evaluated using three main measures: the *Sharpe Ratio* (the most widely used metric to compute portfolio performance), *Value at Risk* and the *Expected Shortfall*. Lastly we will compute an adjusted *Sharpe ratio* using statistical tests and proposed models for corrections (*Bonferroni* method), to give a more realistic view of what out-of-sample portfolio performance could actually look like.

2 Time series generation

Data to test the different trading models on will be generated artificially using the following linear time series function:

$$\Delta y_t - d = \phi(\Delta y_{t-1} - d) + \epsilon_t + \theta\epsilon_{t-1}$$

where $t \in [0, 2000]$, $y_0 = y_1 = 100$, $\phi = 0.6$, $d = 0.025$, $\theta = -0.4$ and $\epsilon \sim N(0, 1)$, which will be generated using *numpy*'s 'default_rng' random number generator with a seed set to my student number (17088101).

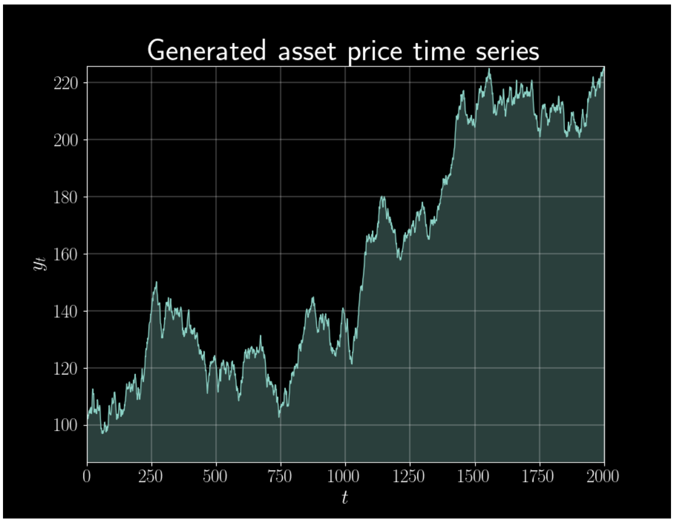

Figure 1: Time series of 2001 generated prices using linear function stated above

The time series will be divided in the following way: the first 1400 observations will be included in the vector of training observations and the following 601 observations will be included in the vector of test observations, to evaluate the specific models with.

3 Methodology

This section will discuss the strategies which will be used to trade on the generated time series, methodology to construct the portfolio and metrics to assess its performance. Furthermore we will discuss the adjustments we will undertake on one of the metrics to give more realistic estimates out-of-sample.

3.1 Trading Strategies

We will define three long-only different trading strategies and a benchmark to compare them with. The strategies will be self-financing (no money will be borrowed and no leverage will be used) and the initial cash in the portfolio C_0 will be 10000. At each time step, the value of the portfolio TV will be updated using the following function:

$$TV(t) = C(t) + p(t)V(t) = C(t+1) + p(t)V(t+1)$$

where TV refers to the total portfolio value, C to the amount of cash held, p to the price of the asset and V to the amount of the asset held. For each time step, logarithmic returns will be computed using the following:

$$r_a(t) = \log\left(\frac{TV_a(t)}{TV_a(t-1)}\right)$$

for each trading strategy a, and each time step t. The parameters of each model will be tuned in the training set and then their performance will be computed in the test set. While this assumes stationarity in the data-set, a condition which is often not held in financial time series, it will ensure the specific models are not overfitted to the underlying data.

3.1.1 Benchmark portfolio (buy-and-hold)

The benchmark portfolio will be computed using a 'buy-and-hold' strategy, where the asset will be bought at the start of the period and sold at the end of the period:

$$TV_t = X_t p_t \;\; \forall t, \quad X_t = X_0 = \frac{C_0}{p_0}, \quad C_t = 0 \;\; \forall t > 0$$

where C_0 is the initial amount of cash and p_0 is the initial price of the asset.

3.1.2 Trend following (moving average cross-over)

Many studies have been conducted highlighting the importance of momentum and trend factors when selecting viable assets to invest in [1] and creating forecasting models to predict future asset price time series. The first trading strategy we will define is a trend-follower based on the cross-over of two simple moving average indicators (one with a short window and one with a long window):

$$\frac{1}{L}\sum_{k=1}^{L} p_{t-k} - \frac{1}{S}\sum_{j=1}^{S} p_{t-j}$$

where S is the window length for the short moving average and L is the window length for the long moving average. If the result of the above formula is negative it means the short moving average is greater than the long moving average and the asset is gaining momentum. We will therefore transfer all of the cash in the portfolio into the asset. If the difference is positive it will signal a loss in momentum, meaning we will liquidate our entire portfolio and hold cash instead.

3.1.3 Mean-reverting (EWMA arbitrage)

The concept of mean-reversion refers to the constant move of a time series towards its mean value. The most classical example is the stochastic process defined as the 'Ornstein-Uhlenbeck' process, originally developed to model the velocity of a brownian particle [2]. In financial theory it is referred to as the 'Vasicek' model [3], used most commonly to describe the behaviour of interest rates:

$$dx_t = \theta(\mu - x_t)dt + \sigma dW_t$$

where μ, θ and σ are all constants, referring to the average value the process is moving towards, the 'strength' of the mean-reversion and the volatility. dW_t is arithmetic brownian motion. In the context of our trading algorithm, we will be modelling the asset price with an exponentially weighted moving average (EWMA):

$$c(\lambda)\sum_{k=1}^{\infty} \lambda^k p_{t-k}$$

where k is the time window of prices considered, which will be tuned in the training set. The algorithm will allocate capital as following: if the price of the asset is below the EWMA, the entirety of the cash currently held will be invested in the asset, as this will signal an arbitrage opportunity (the asset is under-priced). If the opposite occurs, the entirety of the portfolio will be liquidated into cash. By definition, a mean-reverting strategy is expected to behave conversely to a trend-following one, therefore we expect almost inverse results between this strategy and the previously defined one (moving average cross-over).

3.1.4 Autoregressive forecast

In the field of time series econometrics, an autoregressive (AR) process assumes a variable depends on a deterministic component, based on a linear combination of past values and a stochastic component. The AR(p) model has the following function:

$$Y_t = \phi_1 Y_{t-1} + \phi_1 Y_{t-2} + ... + \phi_p Y_{t-p} + \epsilon_t$$

where Y_t is the time series value at time t, ϕ is a scaling applied to past values in the time period and ϵ_t is a stochastic (random) component, which we are going to model as Gaussian white noise. To forecast the values in our time series, we are going to use an AR(1) model, which has the following function:

$$Y_t = \phi_1 Y_{t-1} + \epsilon_t$$

The time series value will only be expected to follow a linear relationship with the previous value in the series. Applying such a model on raw price data will not work very well, this is because we are not looking to forecast a time series which is non-stationary, with mean and variance not constant through time. In order to combat this we will create a new time series out of differences between two steps in the original data ($Y_t - Y_{t-1}$). Stationarity in the two time series will be tested through the 'Augmented Dickey-Fuller Test', which returns a p-value through which we can decide to reject the null-hypothesis of non-stationarity at a certain confidence level. The autoregressive model will then be trained on an established time window, which we will tune in the training set and will be used to forecast the asset price difference of the next time step. All of the forecast differences will then be aggregated into a new time series, which will then be used to create a forecast price series as following:

$$Y_{t+1}^f = Y_t + \Delta Y_t^f$$

where Y_{t+1}^f is the forecast price at time $t+1$, Y_t is the actual price at time t and ΔY_t^f is the forecasted difference at time t.

Following this forecast, we will trade based on momentum (similar to the first one), when the actual price is greater than the predicted one ($Y_t^f > Y_t$), we will allocate all of the cash in the portfolio into the asset. When the asset price is below the predicted one, we will liquidate our portfolio. The reasoning behind this is model is that when the market price is above our prediction, it will gain momentum and is likely to increase in value even more.

3.2 Performance indicators

Our trading strategies will be evaluated using three of the main performance indicators: the *Sharpe ratio*, *Value at Risk* (MDD) and the *Expected Shortfall*. The *Sharpe ratio*, originally introduced by Nobel laureate William Sharpe, is arguably the commonly used metric to evaluate the performance of investment portfolios and is computed as following:

$$SR = \frac{r_t}{\sigma_t}$$

where r_t is the average excess return (return above the risk-free rate r_f) over the time period. For the purpose of this study, we will assume the risk-free rate r_f is 0. The *Sharpe ratio* computed will be annualised (we will multiply this by $\sqrt{252}$, assuming a year has 252 trading days).

Value at Risk estimates how much is the maximum loss that can be experienced in a given time period (which we will pick as one day) at a given confidence level (in our report this will be expressed as a negative value, showing the maximum daily loss). This will be computed by arranging the returns of a given strategy in ascending order and finding the return which is equivalent to the percentile of $(1 - \text{confidence level})$. In the case of our report, we will select the returns at the 5^{th} percentile, as we are aiming to find the *Value at Risk* at a 95% confidence level. This approach is positive as it does not assume any particular distribution of the returns, differently from the variance-covariance approach to compute *Value at Risk*, which draws returns from a normal distribution. We know this characteristic to be wrong due to the statistical properties of asset returns, such as fat-tails.

Although it is widely used in the financial sector, many argue that *Value at Risk* computed on historical returns is an inappropriate measure to use as it assumes the past performance of an asset is indicative of future behaviour, a fact we know does not hold due to the non-stationarity of asset returns. Therefore, as a third performance indicator we will compute *Expected Shortfall*, which is the expected return of a portfolio in the worst $p\%$ of cases, where $p = (1 - \text{confidence level})$. This will be computed by taking the average of all of the returns below the *Value at Risk* at a given confidence level. In our report this will be expressed as a positive value, although it is representing a loss.

3.3 Adjustments and validity of the Sharpe ratio

The estimates for the *Sharpe* ratio we will compute for each portfolio will almost never be reflective of the actual out-of-sample performance of these strategies. This is because of several factors, such data snooping, p-hacking and over-fitting. Data snooping and over-fitting refer to the author of a model tuning it on out-of-sample data, while they should only be estimating such parameters in-sample. This could lead to greatly over-estimated performance of a model, which would then fail when exposed to new data.

P-hacking refers to the process of testing a large amount of models and finding statistically significant results 'by chance' and not because the underlying model is performing well. This leads to false-positives arising, which in the scope of successful trading strategies would mean finding a strategy which is excellent when back-tested but fails when used on unseen data.

We will overcome data snooping bias by only selecting optimal parameters for our models in-sample (training set). To combat p-hacking we will be computing adjusted *Sharpe* ratios with the suggested 'Bonferroni adjustment' [4]. *Sharpe* ratios can be transformed into t-statistics by multiplying them by $\sqrt{\frac{T}{252}}$, where T is the number of samples in the data-set (we are dividing by 252 as we previously annualised the ratio). We will then compute p-values from these t-statistics in order to test the null-hypothesis of a zero *Sharpe* ratio. The Bonferroni adjustment takes into account the number of experiments (trading strategies in our case) considered, M and multiplies each p-value with this. In our analysis $M = 3$, as we are considering 3 trading strategies. After computing this adjustment, we will convert the p-values back into *Sharpe* ratios.

4 Results

4.1 Trend-following

The trading strategy was tuned with different short window lengths $w_s \in [3, 14]$ and long window lengths $w_l \in [6, 28]$ (the models were only tested if the short window $<$ long window). The optimal model (with the highest cumulative return) on the training data was one with a short window $w_s = 3$ and a long window $w_l = 6$, producing cumulative logarithmic returns of 199.54% over the time period against the benchmark of 78.64%. Out-of-sample, the strategy also outperformed the benchmark,

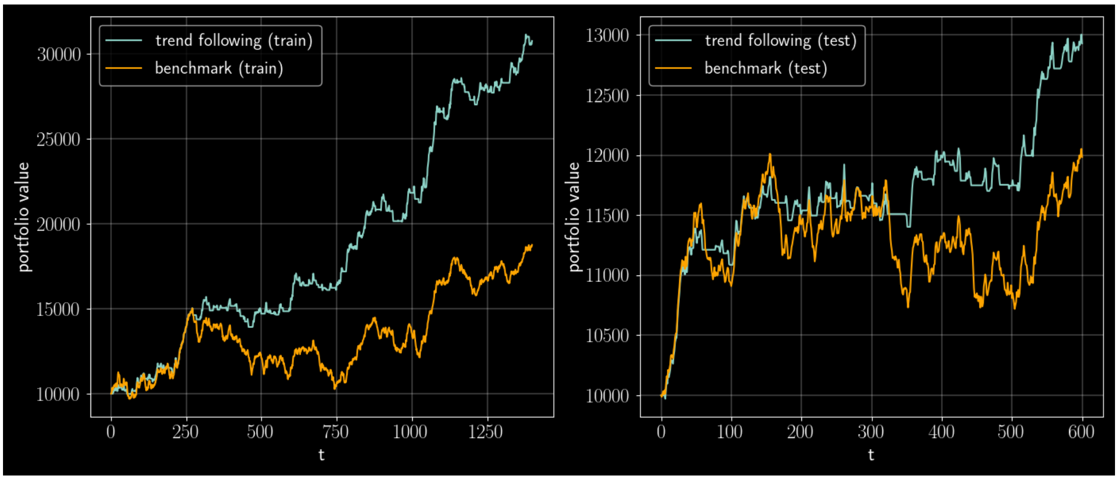

Figure 2: Comparison of performance of trend following strategy on the time series in-sample and out-of-sample

but less drastically, producing cumulative logarithmic returns of 28.71% against the benchmark of 18.93%. This could have been due to the time window out-of-sample being 600 time steps as opposed to the 1400 of the training set, giving the strategy less time to be profitable.

From the table below (Table 1), we can clearly see the trend-following strategy we created based on moving average cross-overs is out-performing the benchmark of buying and holding the asset with all of the metrics.

Trend-Following (out-of-sample)	Benchmark (out-of-sample)
Key performance indicators: *Sharpe ratio* (annualised) $= 1.76$ *Value at Risk* (95% level) $= -0.61\%$ *Expected Shortfall* (95% level) $= 12.01\%$	*Key performance indicators*: *Sharpe ratio* (annualised) $= 0.95$ *Value at Risk* (95% level) $= -0.79\%$ *Expected Shortfall* (95% level) $= 15.99\%$

Table 1: Table showing the main performance indicators for the trend-following strategy out-of-sample compared against the benchmark of a buy-and-hold strategy

4.2 Mean-reverting

The mean-reverting strategy was tuned with different window lengths of the exponentially weighted moving average (EWMA), with different window lengths $w \in [10, 19]$. All of the time windows tested produced a loss in the training set, however the minimal loss was produced with a window of length 10. For comparability purposes, we still computed test set performance, although in a real-life setting the strategy would have likely been scrapped after producing such a loss in-sample.

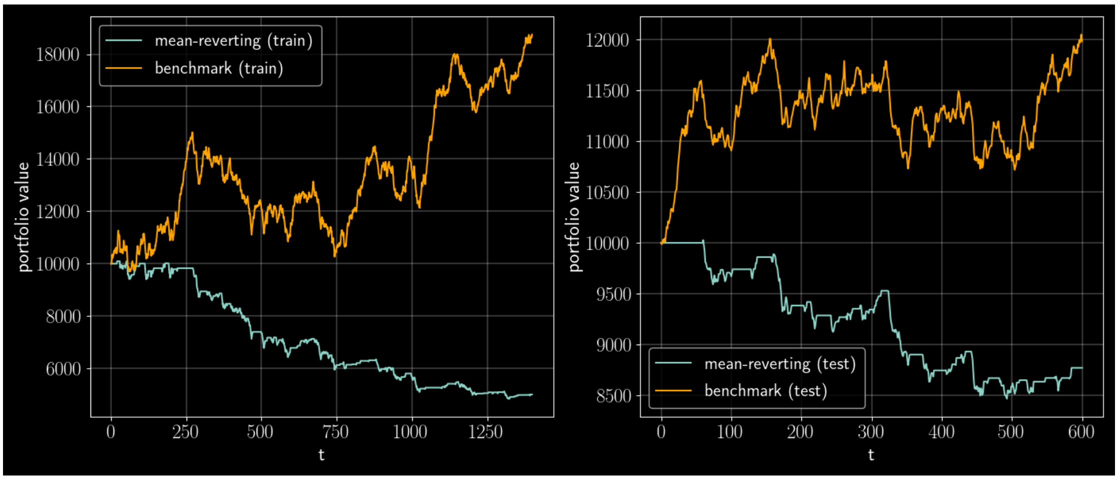

Figure 3: Comparison of performance of mean-reverting strategy on the time series in-sample and out-of-sample

Mean-reverting (out-of-sample)	Benchmark (out-of-sample)
Key performance indicators: *Sharpe ratio* (annualised) $= -1.12$ *Value at Risk* (95% level) $= -0.66\%$ *Expected Shortfall* (95% level) $= 10.56\%$	*Key performance indicators*: *Sharpe ratio* (annualised) $= 0.95$ *Value at Risk* (95% level) $= -0.79\%$ *Expected Shortfall* (95% level) $= 15.99\%$

Table 2: Table showing the main performance indicators for the trend-following strategy out-of-sample compared against the benchmark of a buy-and-hold strategy

From the results outlined in Table 2 we can see our strategy performed very badly on the underlying data, with a *Sharpe ratio* below 0. This poor performance is expected due to the success of the previously tested trend-following strategy. Furthermore, mean-reversion strategies are often successful when applied to a time series which is stationary, a property not possessed by the majority of assets. A potentially better strategy could have been trading on pairs of co-integrated time series [5], the differences of which are more stationary.

4.3 Autoregressive

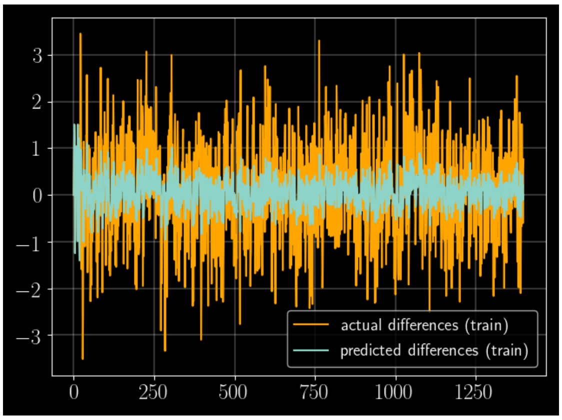

Figure 4: Time series of actual price differences and predicted price differences

Before the implementation of a trading strategy based on an autoregressive forecast, we tested the price data for stationarity using the 'Augmented Dickey-Fuller' test both on the raw price time series generated and on first differences (as defined in section 3). The null-hypothesis of non-stationarity could not be rejected for the price data at the 95% confidence level with a p-value of 0.8813. On the other hand, we could reject the null hypothesis of non-stationarity on the first differences of the data at the 95% confidence level with a p-value of 0.0, therefore we trained the AR(1) model and forecasted the differenced data (Figure 4).

The autoregressive model was tuned in the train set with 5 different window lengths $w = [10, 20, 50, 100, 200]$, where the optimal result was achieved with a window $w = 20$. The strategy achieved cumulative logarithmic returns of 295.67% in-sample against the benchmark of 78.64% and 32.37% out-of-sample against the benchmark of 18.93%.

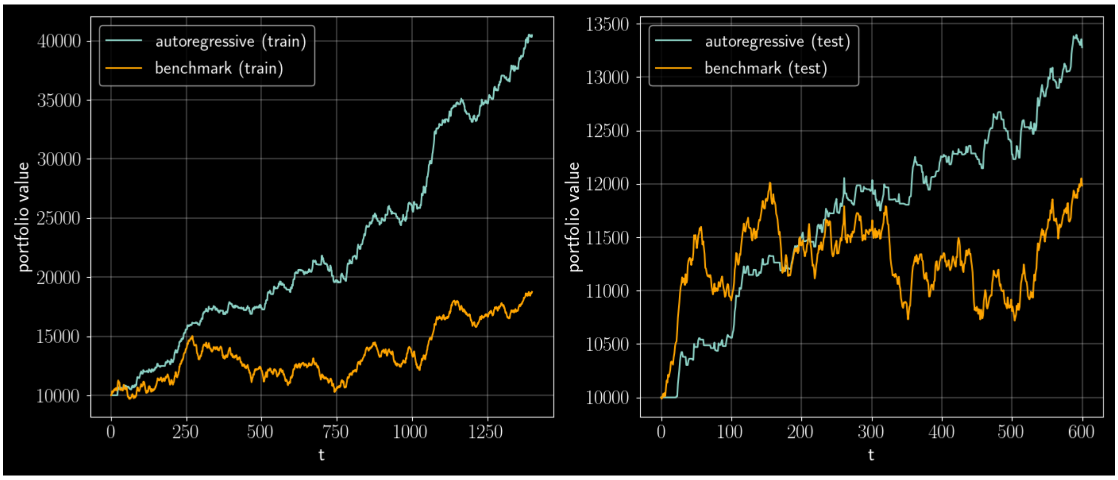

Figure 5: Comparison of performance of autoregressive strategy on the time series in-sample and out-of-sample

Autoregressive (out-of-sample)	Benchmark (out-of-sample)
Key performance indicators: *Sharpe ratio* (annualised) $= 2.31$ *Value at Risk* (95% level) $= -0.52\%$ *Expected Shortfall* (95% level) $= 9.91\%$	*Key performance indicators*: *Sharpe ratio* (annualised) $= 0.95$ *Value at Risk* (95% level) $= -0.79\%$ *Expected Shortfall* (95% level) $= 15.99\%$

Table 3: Table showing the main performance indicators for the trend-following strategy out-of-sample compared against the benchmark of a buy-and-hold strategy

4.4 Analysis

Comparing the performance of the three trading strategies against the benchmark we can clearly see the trend-following and autoregressive techniques are out-performing the benchmark portfolio of buying and holding the asset, while the mean-reverting strategy is losing money in both cases. The autoregressive strategy is out-performing the others not only in terms of cumulative returns, but also when adjusting for risk, having the lowest *Value at Risk* and *Expected Shortfall*.

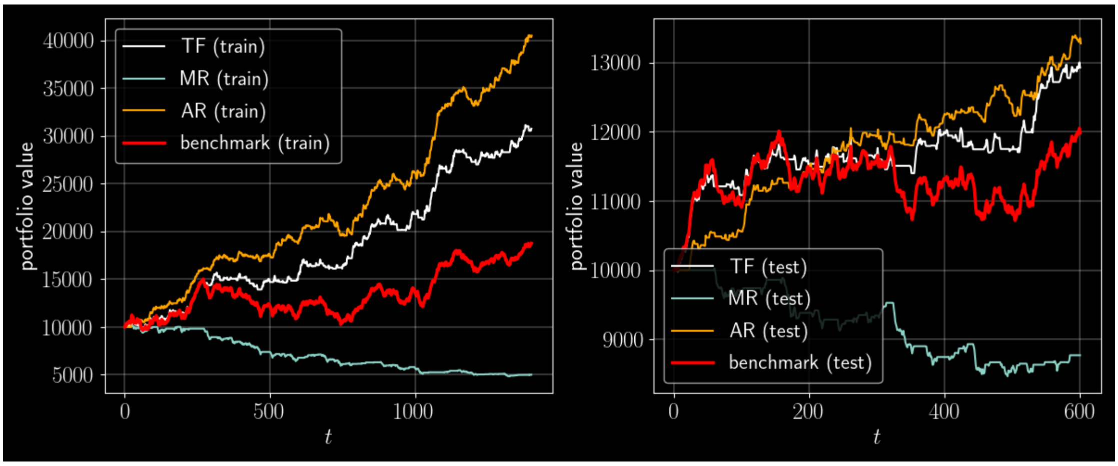

Figure 6: Compared performance of benchmark portfolio, trend-following (TF), mean-reverting (MR) and autoregressive (AR) strategies

This could have been due to the function generating the data[1] being an ARIMA(1,1,1), which contains the AR(1) component we are using to forecast values. In a real setting, we would be fitting trading strategies with empirical data, so this model could potentially perform worse. Results could have increased using a more sophisticated time series forecasting technique, such as an ARMA model (to incorporate a moving average component as well) or more modern techniques in machine learning, such as LSTM networks[2]. Furthermore, as previously mentioned in this report, the mean-reverting strategy could have been implemented on pairs of co-integrated assets with stationary differences, to achieve better cumulative returns.

Another fault in the models proposed is the lack of inclusion of transaction costs, which can greatly

[1]The function we generated our time series with is $\Delta y_t - d = \phi(\Delta y_{t-1} - d) + \epsilon_t + \theta\epsilon_{t-1}$

[2]Long short-term memory networks

reduce the performance of a strategy. While both the trend-following and autoregressive trading models are out-performing our benchmark, they are trading very frequently (AR is trading almost at every time step). Taking into account transaction costs would likely lead to lower returns and a more realistic portfolio performance.

Lastly, although we are accounting for risk in our performance metrics, we are not implementing any risk-measures when executing the trading strategies. When trades are placed, they allocate 100% of the investor's portfolio into the asset or liquidate 100% of the assets into cash, a methodology which would likely not be used in a real setting.

4.5 Validity of the Sharpe ratio

For each of the trading strategies proposed (with *Sharpe ratio* > 0), we computed the adjusted *Sharpe ratio* using the *Bonferroni* adjustment (mentioned in the Section 3.3).

Original values	**Adjusted values**
Sharpe p-values: Trend-following $= 0.0828$ Autoregressive $= 0.0587$	*Adjusted Sharpe p-values*: Trend-following $= 0.2484$ Autoregressive $= 0.1762$

Table 4: P-values to test the null-hypothesis of a non-zero *Sharpe ratio* before and after *Bonferroni* adjustment

From Table 4 we can clearly see the *Sharpe ratios* are not statistically significant after the *Bonferroni* adjustment. The p-values are above 0.05, therefore we cannot reject the null hypothesis of a non-zero *Sharpe ratio* at the 95% confidence level. Transforming the p-values back into *Sharpe ratios*, we obtain 0.5832 for the trend-following strategy and 0.8840 for the autoregressive strategies. While these are much lower than the original measures, they are likely to be more reflective of the actual performance of such strategies when exposed with new data. However, the *Bonferroni* adjustment is considered a very harsh adjustment, often more appliccable to the 'hard sciences', rather than the field of economics and finance, which could lead to the *Sharpe ratios* actually being underestimated.

5 Conclusion

In conclusion, this report outlined the methodology and results of three defined trading strategies on an artificially generated time series. The trend-following and autoregressive strategy greatly outperformed both the benchmark and the mean-reverting strategies in terms of cumulative logarithmic returns, *Sharpe ratios* and risk measures. This could have been due to the structure of the data, generated from an ARIMA(1,1,1) process. The strategies would need to be adapted before the implementation in a real setting, taking into account other factors, such as the existence of transaction costs or a non-zero risk-free rate r_f.

References

[1] Lemprire Y, Deremble C, Seager P, Potters M, Bouchaud JP. Two centuries of trend following. The Journal of Investment Strategies. 2014;3(3):41–61.

[2] Uhlenbeck GE, Ornstein LS. On the Theory of the Brownian Motion. Physical Review. 1930;36(5):823–841.

[3] Bjork T. Arbitrage theory in continuous time. Oxford University Press; 1998.

[4] Harvey CR, Liu Y. Backtesting. The Journal of Portfolio Management. 2015;42(1):13–28.

[5] Elliott RJ, Van Der Hoek * J, Malcolm WP. Pairs trading. Quantitative Finance. 2005;5(3):271–276.