Short-term forecasting of non-stationary time series using multiple feature selection methods

Muhammed Imran Özyar[1]

[1]Delft University of Technology

***Abstract*—Time series forecasting has been proved to be relatively easier for stationary time series, compared to non-stationary time series. This research proposes a method to partially omit the non-stationarity of the data using prioritized sampling. Using multiple feature selection methods in combination with a random forest regressor (RFR), we aim to predict the values for a non-stationary time series. In particular, the principal component analysis (PCA), kernel PCA, incremental PCA and independent component analysis methods are used. The features extracted from these methods will be fed into an RFR both individually and combined, using the union and intersection operators. The features given by the IPCA ∪ PCA ∪ KPCA method, using prioritized sampling with multiple features per day provide the best improvement over the baseline.**

***Index Terms*—time series, forecasting, non-stationary, feature selection method, prioritized sampling**

I. Introduction

Forecasting for a series of observations of a variable can be a cumbersome task. If the observations of a variable are made over time, then the series is called a *time series*. Examples of time series are given by the stock market, where the close prices of a stock are observed over time, and by traffic, where e.g. the amount of vehicles passing by a single point are observed over time. The prediction of such time series gets easier when the observed values tend to show the same statistical properties, independent of which time the measurements are made at. This time series, even though the preceding definition was roughly put, is defined as (weakly) stationary. For a stationary time series up until time t we can simply assume that in step $t+1$, the distribution of the time series will remain the same and we can make predictions based on this assumption.

However, there still remain many non-stationary time series that do not have this property (e.g. in economics or sounds analysis) [6] and thus, cannot be predicted using traditional forecasting techniques, since these techniques, such as ARIMA, rely on the stationary assumption [4]. In their research on short term traffic flow forecasting, [25] note that predictions of traffic flow time series do not give any other information than the prediction itself. This forms a barrier for traffic operations and management, since they will not get to know why certain predictions are given and thus cannot fully act upon them. However, this observation is not limited to traffic flow forecasting alone. Additional information in forecasting time series in any domain can be, although domain specific, beneficial. A step towards adding information to the prediction can be made through the use of feature selection, where the features can be researched for additional information.

The hypothesis made in this research is that some non-stationary time series contain stationary components. These components should be captured using prioritized sampling, where the stationary components are sampled with a higher probability than the non-stationary ones. These samples can then be used for the prediction. A suiting sampling method is proposed in Section IV-A.

The goal of this research is to construct a method for the short-term forecasting of non-stationary time series using multiple feature selection methods. To do this, the stationary parts have to be extracted from the time series using a sampling method, while also taking the non-stationary parts in consideration when making the prediction. In this step, we aim to partially omit the non-stationarity of the time series, such that our training data will be biased towards the stationary ones (i). It is furthermore noted by [14] that prediction accuracy can be improved using an additional variable that is closely correlated with the current one. This might be present in the current data set, but can also be from another one. Thereafter, its usefulness will be analyzed using correlation analysis (ii). Finally, using dimensionality reduction methods, namely principal component analysis (PCA), incremental PCA (IPCA), kernel PCA (KPCA) and independent component analysis (ICA), as feature selection methods and therefore as the input for a random forest regressor (RFR), a prediction will be given for the time series using the additional variable (iii). In their paper on multiple feature selection methods, [24] have indicated to use other methods of feature selection.

The novel approaches are the applications of the KPCA, IPCA and ICA methods. The first baseline that results will be compared to are the predictions given by an RFR, without using any feature selection methods. The PCA method has been included to provide an additional baseline to compare the other results to. The acquired results will be compared to the baselines and to each other.

II. Related work

Tsai et al. [24] have, similar to this research, used a combination of multiple feature selection methods and an artificial neural network (ANN) to forecast economical time series. One of the differences of this research with that of [24], is that they had used a classifier instead of a regressor. By using a classifier, they predicted the up or downward movement of the time series. In this research, however, a random forest regressor will be used to predict the actual time series. The choice for a regressor is preferred since it adds more information to the prediction, while the prediction of an up or downward movement does not include information of the intensity of the movement. Because of the difference in classification and regression, the accuracy metric used by [24] cannot be used. Instead, the mean squared error will be used.

Furthermore, where [24] have used a sliding window to determine the train and test sets, we have used a prioritized sampling method, after which the acquired samples are randomly split into the train and test sets of sizes $\frac{2}{3}$ and $\frac{1}{3}$ of the samples, respectively.

Additionally, even though [24] have achieved accurate results for their data sets using feature selection and an ANN, the essence of feature selection, namely the extraction of useful features for data representation and potential explainability of generated results, gets overshadowed by the complexity and lack of interpretability of the ANN. Hence, the usage of a random forest becomes apparent.

III. Theoretical Background

In this section, the required knowledge for the running example and the used methods is explained. Namely, time series are discussed in more detail and two stationarity tests for time series are explained. Additionally, the variable that will be analyzed and predicted on is explained, along with the feature selection methods that will be used. Finally, the workings of the random forest regressor are explained and the evaluation metrics are discussed.

A. Time series

A time series, denoted as $\{X_t\}$, is a set of observations x_t at time t. *Stationary* time series are time series whose, roughly put, statistical properties are independent of the time axis. To understand exactly what stationary time series are, we use the following definitions: $\mu_X(t) = \mathbb{E}[X_t]$ and $\gamma_X(r, s) = \text{Cov}(X_r, X_s)$. Using this terminology, (weakly) stationary time series are then defined by Definition 1 for a series $\{X_t\}$ [4].

Definition 1. ***Stationary Time Series***
$\mu_x(t)$ is independent of t and $\gamma_X(t, t+h)$ is independent of t for each h

It is also noted by [4] that some time series that may seem non-stationary can be made stationary by removing trends through mathematical operations, which will not be discussed in this paper. Not all time series, however, have the property of removable trends. Therefore, in this research we will assume the negation of Definition 1. We define this assumption using Definition 2.

Definition 2. ***Non-stationary Time Series***
$\mu_x(t)$ is dependent on t or $\gamma_X(t, t+h)$ is dependent on t for at least one h.

The stationarity of a time series can be tested with statistical tests. Two of those will be used to check for stationarity in our data, namely the Augmented Dickey-Fuller (ADF) test and the Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test. Both tests check for unit roots in the time series. The ADF test is used for testing the null hypothesis, being that the time series has a unit root, meaning it is non-stationary. The alternative hypothesis in this case is that the time series does not have a unit root and is thus stationary [7]. The KPSS test is also used for testing the null hypothesis, being that the time series does not have a unit root and is thus stationary. The alternative hypothesis is consequently that the time series does have a unit root and is non-stationary [13]. For both tests, a significance of $\alpha = 0.05$ will be used, meaning that if the resulting p-values are smaller than or equal to 0.05, we will reject the null hypothesis. If this is not the case, thus the p-values are larger than 0.05, then we fail to reject the null-hypothesis.

B. Time series variable

As explained in Section III-A, a time series is a series of observations for a variable X. This paper will use historical stock prices as the running example. It may be intuitive to think that the variable chosen for the time series would be the close prices of a stock in

every day, however, this is not preferred. This is because stock prices can be too volatile for statistical analysis. Instead, the *logarithmic returns* are used. To understand this concept, first, *simple returns* should be introduced. Simple returns are defined as the rate of change between day t and $t-1$, which is mathematically defined as

$$sr_t = \frac{s_t - s_{t-1}}{s_{t-1}},$$

for a stock price s_t at day t. Simple returns can be seen as the first order Taylor expansion of the log of one plus the simple returns, such that $\log(1 + sr_t) = sr_t$, for arbitrarily small sr_t, such that $sr_t^2 < \epsilon$ for $\epsilon \to 0$. Therefore, logarithmic returns are defined as

$$\log(1 + sr_t) = \log\left(1 + \frac{s_t - s_{t-1}}{s_{t-1}}\right) = \log\left(\frac{s_t}{s_{t-1}}\right),$$

for a stock price s_t and simple returns sr_t at day t. As mentioned by [8], logarithmic returns are time additive, and it is easier to derive the time series properties of additive processes than multiplicative processes. In this context, [8] notes that estimating returns over longer periods using the simple returns can be quite unsatisfactory. Because of this, we will be using the logarithmic returns as the observed variable. Logarithmic returns will be referred to as log returns further on.

C. Autocorrelation function

The autocorrelation function, is defined as the correlation of a variable with an earlier occurrence of the variable. In the context of time series, the autocorrelation function denotes the correlation of a variable X_t and a variable X_{t+h} for a lag $h \in \mathbb{Z}$. The lag thus denotes the difference in measurement on the time axis. First, in the terminology introduced in Section III-A, let us define $\gamma_X(h) = \gamma_X(t+h, t)$. Mathematically, [4] defines the autocorrelation function then as

$$\rho_X(h) = \frac{\gamma_X(h)}{\gamma_X(0)} = \text{Cor}(X_{t+h}, X_t).$$

The autocorrelation function proves to be useful in determining the batch size when sampling data from the time series. The batch size will be defined in terms of lag h, such that the autocorrelation function until that specific lag of h emits a sufficiently large correlation. The size of the correlation is evaluated against a $100 - \alpha\%$ confidence interval.

D. Feature selection

1) Principal component analysis: Principal component analysis (PCA) is a statistical procedure that, using mathematical transformations, converts a set of possibly correlated variables into a set of linearly uncorrelated variables, called principal components [10]. Each principal component lies on a corresponding principal axis. This process is called dimensionality reduction and can be used as a feature selection method, as done by [24]. The amount of principal components extracted from the PCA is less than or equal to the amount of initial variables. The exact amount of principal components to be extracted from the PCA method can be acquired through a cumulative explained variance plot against the amount of components. This plot show the amount of variance that is explained by an n amount of principal components. The explained variance is preferred to lie close to 1, such that almost all of the variance within the data set is covered.

2) Incremental principal component analysis: Incremental principal component analysis (IPCA) is similar to PCA. IPCA is more memory-efficient than PCA [16], but there is no guarantee on the approximation of the principal components [26]. In the specific example given in this paper, PCA poses no problem to the memory, since the size of the input data is not large. However, this does not mean that this will be the case for other areas where this method can be applied. Therefore, IPCA is also covered in this paper.

3) Kernel principal component analysis: Kernel principal component analysis (KPCA) is an extension of PCA, in a sense that KPCA enables non-linear dimensionality reductions [21]. This method is used for the comparison of results with PCA, as mentioned by [24].

4) Independent component analysis: Independent component analysis (ICA) can be seen as an extension of PCA. ICA consists of searching for a linear transformation of the initial non-Gaussian data, such that the statistical dependence between components is minimized [5]. It is noted by [9] that such a construction seems to capture the essence of the data in the field of feature selection and is therefore also used in this research.

E. Random forest

To understand random forests, first classification and regression trees (CARTs) have to be introduced. CARTs first partition the training data into different regions. These regions are accessed through binary conditions, given by the *splitting variables*. For any new input to make a classification or regression on, the region the input falls in has to be determined. This is done through following a path though the splitting variables, down to the leaf nodes. The average value of the data points in the region the new input data falls into is the prediction in case of regression. The structure of a regression tree

is determined by finding the right amount of regions, splitting variables and their corresponding thresholds through the minimization of the MSE [23].

An advantage of CARTs are that no inherent assumptions are made on the distribution of the input data. Another advantage is that CARTs are relatively easy to interpret for non-statisticians, in a sense that the decisions made by the tree are understandable and explainable [15].

Random forests use multiple CARTs in combination with bagging, with an additional layer. In bagging, successive CARTs do not depend on earlier CARTs, thus every CART is independently constructed [2]. The additional layer changes the way CARTs are constructed, such that every node is split using the best among a subset of predictors randomly chosen by that tree [3]. Random forests using regression trees are often referred to as regression trees and usually take the weighted or unweighted average of the trees in the forest [22].

Random forests have the same property as CARTs in a sense that the produced results are easy to interpret for non-statisticians. Additionally, random forests are robust against overfitting [3].

F. Evaluation metric

As opposed to [24], this research is considered with prediction in the form of a regression for time series. A widely used metric for measuring the quality of an estimator is the mean squared error (MSE), which is defined as

$$\text{MSE} = \frac{1}{N}\sum_{i=0}^{N}(X_i - \hat{X}_i)^2,$$

for a variable $X_i \in X$, a prediction $\hat{X}_i \in \hat{X}$ and $N = |X|$. As can be seen, this metric measures the average squared error of an estimator. This implies that the closer the MSE lies to zero the smaller the error is and thus, the better the estimator is.

Additionally, when combining feature selection methods, different amount of features will be selected. It is expected to have lower MSEs for combinations having many features, since these features would act as additional explanatory variables [11]. To counter this effect, we introduce another measurement, which weights the MSE based on the amount of features, relatively. This method of evaluating models is preferred in cases where the amount of features are critical in the selection of a model. This metric gives a relative score that can be used to compare the used models. We call this the feature weighted MSE score (FWMS) and it is defined as

$$\text{FWMS}(t) = \left(\frac{F}{\hat{F}}\right)\left(\frac{M}{\hat{M}}\right)^t, \tag{1}$$

for an MSE of M and F features of a feature selection method. The t variable indicates the spread of the relative MSE. Furthermore, the $\hat{M}$ and $\hat{F}$ variables are then defined as the averages of M and F, respectively, over all feature selection methods. Ideally, we want $F < \hat{F}$ and $M < \hat{M}$, since this would mean that the model achieves a relatively low MSE using relatively few features, as FWMS $\to 0$. The MSE spread variable t determines how large the spread of $\frac{M}{\hat{M}}$ will be. For $t \to 0$, $\frac{M}{\hat{M}}$ will revolve closely around 1 for $|M - \hat{M}| < \epsilon$, $\epsilon \to 0$. However, for $t \to \infty$, $\frac{M}{\hat{M}}$ will spread further away from 1, even for $|M - \hat{M}| < \epsilon$. Because of this effect, Eq. 1 will assume values in a larger range.

IV. Method

This section discusses the methods used for the research, including the sampling method (i), the selection of an additional variable and correlation analysis (ii) and the feature selection algorithm along with the random forest regressor (iii). The sampling method involves prioritized sampling, which, as the name suggests, prioritizes certain regions in the data set to sample from over other regions. Furthermore, an additional variable is selected based on suggestions from literature and its correlation with our time series is analyzed. Finally, the feature selection methods are evaluated individually and combined, using the union and intersection operators, using a random forest regressor.

A. Sampling

Using a suitable sampling method, we want to sample the stationary components from the time series with a higher probability than the non-stationary component. We will call this method *prioritized sampling*, named after prioritizing stationary components over non-stationary components. We will sample a batch of values, instead of sampling a single value, such that the batch, representing a sequence of days, can be used as the input for the models.

The data used in this research is acquired from the open data sets on Kaggle, which provides historical data of the NASDAQ, NYSE and AMEX stock markets. Let us look at the log returns of CPS Technologies (CPSH) from 2011 until 2019 (Figure 1). It can be seen that in the domain of 2012 - 2014, the log returns had a different range of occurrence, namely in $[-0.4, 0.4]$. A simple solution would be to simply cut off the data

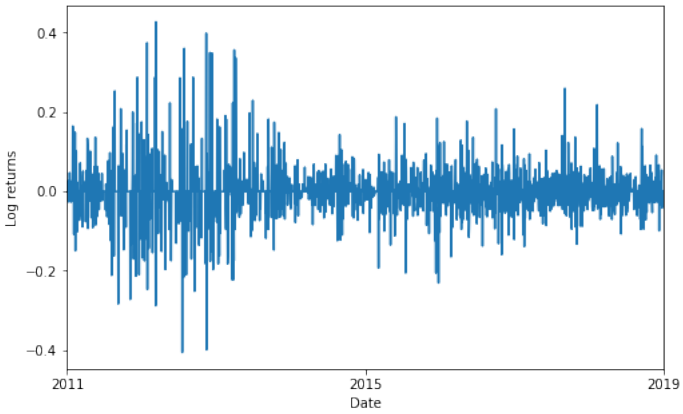

Fig. 1. Log returns for CPSH from 2011 until 2019.

until 2014 and sample the remaining data using uniform, sequential sampling. However, by doing this, we may lose the information of such an event reoccurring and our predictions could be inaccurate. Furthermore, we also see that the log returns after 2014 (and before 2012) seem to be stationary. This observation is confirmed by the results acquired in Section V-A, where the ADF and KPSS tests are deployed. Because of this behavior, this time series can be used as an example. Although drawing conclusions from an observation in the data may work, automating this process or having a systematic method for it is not provided. The field of extreme value theory may provide methods, such as the Peak-Over-Threshold (POT) method, to systematically assess non-stationary regions.

In essence, by using prioritized sampling, we want to bias our samples towards stationary values, such that we help our model in reproducing these regions. Let P denote the probability of sampling a stationary batch. Then we define $P = a * Q$, where a is a positive number and Q is the probability of sampling a non-stationary batch. Since $\sum_{\omega \in \Omega} p(\omega) = 1$, for a probability space Ω and probability mass function $p(x)$ [12], we have $P + Q = 1$. Filling in and solving for Q gives $Q = \frac{1}{a+1}$. Using this formula, we can make sure that we sample one non-stationary batch for every a stationary batches.

It is, however, not preferable to have Q larger than the fraction of the non-sequential values with respect to the whole data set. In this case, this fraction equals to $\frac{1}{3.75}$. If this probability for Q would be used, then we would acquire the same samples as we would with sequential sampling. This is because in sequential sampling, $\frac{1}{3.75}$ of the acquired samples consists of non-stationary values. Sequential sampled data is simply a loop through the whole data, in chronological order, adding batches of a given size z, such that data in the form of $f_i, ..., f_i + z$ is acquired for a feature $f_i, i = 0, 1, ..., N - z$, where N is the size of the data set.

Therefore, we need $\frac{1}{a+1} < \frac{1}{3.75}$. Solving for a results in $a > 2.75$. A plot for various values of a is shown in Figure 2. Note that for $a \to \infty$, $\lim_{a \to \infty} \frac{1}{a+1}$ approaches 0, thus the probability of taking a sample from the non-stationary region is very close to zero. The anticipated effect for $a \to \infty$ are that the predictions become biased towards the stationary values since only samples from those regions are drawn. Because of this, a higher MSE is expected for this value compared to MSEs for $a \neq \infty$.

The amount of samples also plays a role in the sampling method. The data contains a total of 2087 entries. Obviously, sampling 100% of the data will not distinguish the prioritized samples from the sequential samples. However, we still want an optimal amount of samples for all of our experiments. Therefore, we will still use 100% of the data for sequential sampling. For $a \to \infty$, this percentage is determined to be 72% of the data, which is roughly the amount of stationary values in the data. This percentage is just a little less than $1 - \frac{1}{3.75} = 73.33$, such that the sampling algorithm does not have to include all of the stationary values, which will make the algorithm finish faster. For $a \neq \infty$, this value has been determined semi-empirically, meaning that we took the average of the previous percentages, being $\frac{72\%+100\%}{2}$, which resulted in 86% of the data and empirically confirmed it to be the optimal value in this case. However, there is as of now no systematic way of determining the amount of samples for these values of a, which further research may provide more clarity on.

A simulation for validation has been done using Algorithm 1. In lines 5 and 7, a batch of a given size from the stationary values and from the non-stationary values, respectively, is added to the sample set S. In the simulation, this batch size equals to 5, but will be determined exactly in Section IV-B. Furthermore, two means are calculated, namely the arithmetic mean of the entire data set and the mean of samples produced by the simulation for $a = 3$. The comparison of the two means is done to validate the results of the simulation and to check whether we can indeed approximate the chosen value for a. Additionally, the mean of the stationary values and non-stationary values, separately, are calculated. This is done such that we can conclude that the observed non-stationary region is actually non-stationary.

B. Correlations

It is noted by [14] that the use of an additional variable may improve prediction accuracy. In the case of the prediction of stocks, [19] suggests the use of trading volume

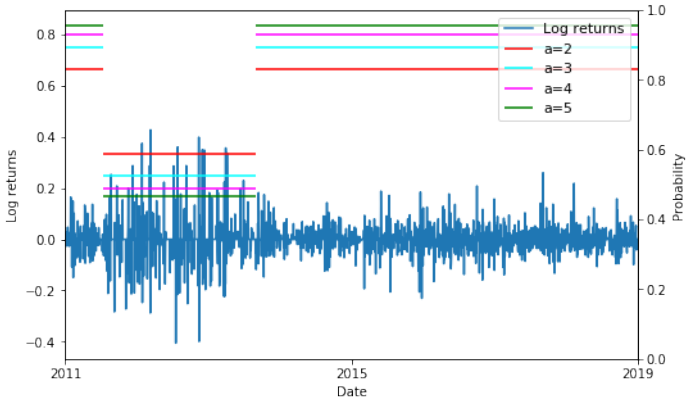

Fig. 2. Log returns for CPSH from 2011 until 2019 with the probability of sampling a batch from the indicated regions for values of $a = 2$, $a = 3$, $a = 4$ and $a = 5$.

Algorithm 1 Pseudocode for the simulation

```
S ← ∅
W ← 50000
while W > 0 do
    if rand ≤ P then
        Add batch to S from stat
    else
        Add batch to S from non-stat
    end if
    W ← W − 1
end while
```

as an additional variable for the prediction. A correlation analysis should be deployed on the trading volume and the log returns to deduce a correlation between the two variables. Pearson's correlation coefficient proposes a measure for linear correlations, as mentioned by [1], but it is not yet known whether the two variables are linearly correlated. Instead, a histogram will be drawn to depict the correlation. This will be done through the creation of a probability distribution for the log returns using the trading volumes. We define this distribution as follows: let L_t be a random variable, representing the log returns on day t. Furthermore, let m be the amount of days that are being analyzed, then for $k = 1, 2, ..., m$, let V be the list of tuples of log returns l_k on day k and trading volumes v_{k-1} on day $k-1$ divided by 1000, such that a list of tuples (l_k, v_{k-1}) is acquired. In practice, a new list is constructed, to which all values l_k are added v_{k-1} amount of times, hence the division by 1000 such that the acquired list will not be too large to prevent long iteration times. Additionally, let $u \in U = \{l \mid (l, v) \in V\}$. We then define

$$\Pr(L_t = u) = \frac{\sum\limits_{(u,v)\in V} v}{\sum\limits_{(l,v)\in V} v}, \tag{2}$$

such that the probability of a log return on day t having a value of u is equal to the fraction of the sum of all trading volumes (divided by 1000) for the log return u divided by the sum of all trading volumes. To furthermore prove that this is a valid probability distribution, it is easy to see that $\sum_{u\in U} Pr(L_t = u) = 1$, since the numerator in Eq. 2 will sum to the denominator in Eq. 2 for all unique values of u:

$$\sum_{u\in U}\left(\sum_{(u,v)\in V} v\right) = \sum_{(l,v)\in V} v$$

Using this probability distribution function, the histogram in Figure 3 is plotted, using 30 bins. For the sake of relevancy, the extreme values below -0.2 and above 0.2 have been cut off. To this histogram, a Student-t distribution with $\nu = 2.56$ has been fitted, using a scale factor of $3.11 * 10^{-2}$ and multiplying the PDF with 400. To check whether the two probability distributions, denoted further as N and M, differ, a Kolmogorov-Smirnov test is applied. This hypothesis test uses the null hypothesis that the distributions are the same and the alternative hypothesis that the distributions differ. A significance level of $\alpha = 0.05$ will be used, such that the acquired test statistic $D_{N,M} > 1.22\sqrt{\frac{|N|+|M|}{|N|*|M|}}$ or the acquired p-value should be smaller than 0.05 for us to reject the null hypothesis [17]. Essentially, this means that $D_{N,M} > 10.77 * 10^{-3}$ or that $p < 0.05$ to reject the null hypothesis. If we are not able to reject the null hypothesis, the underlying distributions are likely to be the same, which means that log returns at day t are Student-t correlated with the trading volumes at day $t-1$. The results of the test are noted in Section V-B.

Additionally, the Pearson correlation for the whole data set is plotted using a heatmap in Figure 4. This heatmap gives the correlation between the volume, open, high, low and close (OHLC) prices, years and the log returns at time t and the same variables at time $t-1$. Since the time series we use is that of the log returns, we are only interested in variables that are correlated with this. Note that the correlation between the OHLC variables is close to 1, since these variables lie close to each other when looked at a daily domain. As previously mentioned, the Pearson correlation for the log returns and the volume is close to zero, which is confirmed by the histogram in Figure 3, since the two variables are not

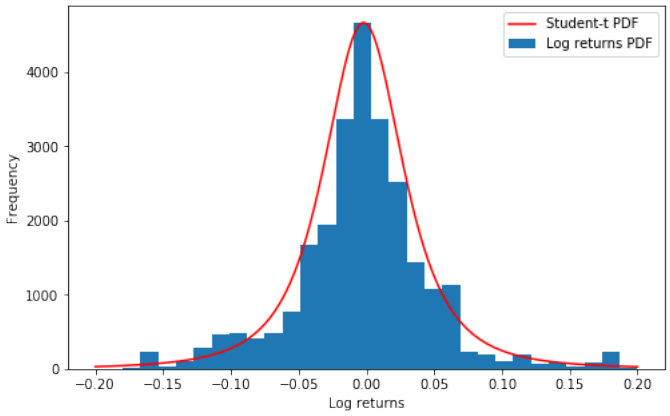

Fig. 3. The log return PDF is depicted in blue with cut off extremes, being smaller than -0.2 or larger than 0.2. The Student-t distribution, along with $\nu = 2.56$, a scale factor of $3.11 * 10^{-2}$ and multiplied by 400, is depicted in red.

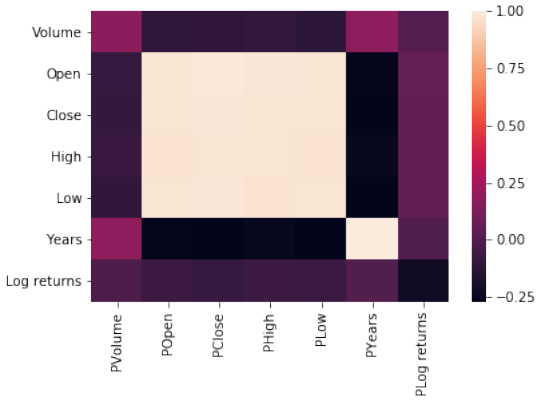

Fig. 4. Pearson correlation between the variables at time t (y-axis) and at time $t-1$ (x-axis).

linearly correlated. The correlation between the OHLC prices and the log returns also seems close to zero.

Furthermore we aim to analyze the autocorrelation based on the log returns from the prioritized samples. In this analysis, the simulation described in Section IV-A is run again with a sufficiently large batch size (20), such that we can analyze the AC function for lags up to 20. The plot is displayed in Figure 5.

From Figure 5, it can be seen that the ACF seems to show relevant correlations until a lag of 9. Because of this, the batch size mentioned in Section IV-A gets corrected from 5 to 10. This is done because the lag at 0 was omitted.

C. Feature selection and regression

In Section IV-B, it was found that the trading volume of CPSH is likely to be Student-t correlated with its log returns. It was furthermore found that the usable batch size for prioritized sampling (PS) for the input of the

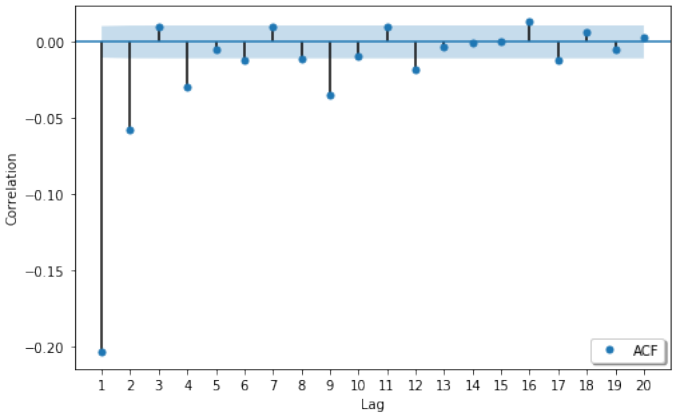

Fig. 5. Plotted AC function for the prioritized CPSH samples (2011-2019). The blue region depicts a confidence interval of 95% ($\alpha = 0.05$). The point for lag = 0 is omitted since this correlation would equal to 1.

feature selection model will equal to 10. This means that there will be used a total of 10 + 10 = 20 variables per training instance, being 10 log returns on sequential days and their corresponding trading volumes. Out of the 20 variables, 18 variables will be used as the data for 9 sequential days. The log return of the 10th day will be used as the target for the prediction, whereas the volume of the 10th day will be dismissed. The 18 variables will be preprocessed (PP), such that every value will lie in the range $[0, 1]$. The PP data will be fed to the feature selection methods, being PCA, IPCA, KPCA and ICA, after which they will both independently and combined be fed to the random forest regressor, from which a mean squared error will result. The preprocessed data will also be fed to the RFR, without feature selection, to set a baseline to compare the results of the feature selection models to. This procedure is depicted in Figure 6. The block depicted by the N, U and I tag, represents the individual evaluation of the feature selection methods (N), the combination of the methods using the union operator (U) and the intersection operator (I).

To determine the amount of components to be extracted from PCA, a plot of the amount of components against the cumulative explained variance can be made. These components will be the features that will be used and will be fed into the RFR. This graph is depicted in Figure 7. It can be seen that at 15 components, more or less all of the variance within the data set is explained. Therefore, 15 features will be used to be fed into the RFR. For comparability, the same amount of features for the IPCA, KPCA and ICA methods will be used.

The KPCA method enables us to use a multitude of kernel methods to recognize patterns within the data. In this study, we use the common Gaussian radial basis

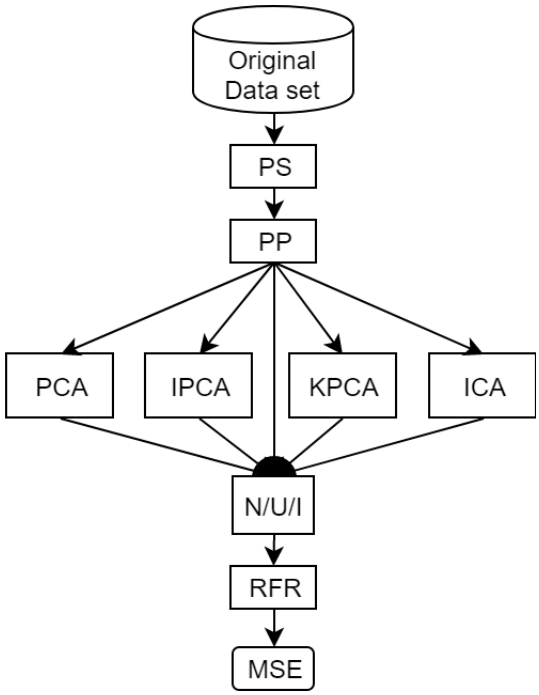

Fig. 6. The procedure of the first and second experiments.

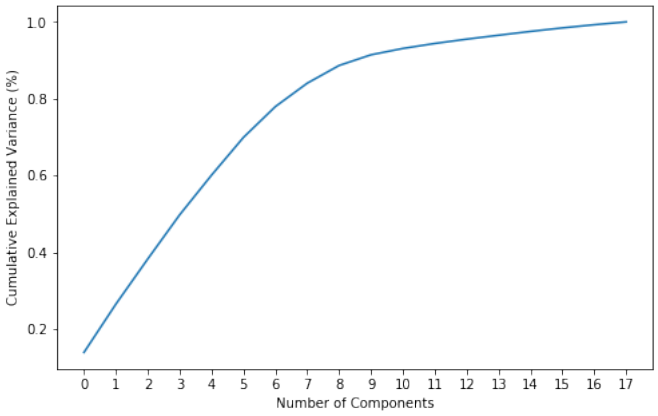

Fig. 7. The amount of components of PCA against the cumulative explained variance.

function (RBF) kernel for KPCA, along with $\gamma = 5$ which was determined empirically. The RBF implementation in KPCA does not yield the principal component axes. Instead, the obtained eigenvectors are seen as projections of the data onto the principal components [20]. For the sake of generality, these eigenvectors shall still be referred to as the principal axes.

The preprocessed data is fed into the feature selection methods with a feature size of 15 and are hereafter fed into an RFR with a depth of 20 and a tree count of 500. The depth has been chosen such that it is slightly bigger than the feature size, such that every tree has the room to process some features more than once. The tree count has been chosen sufficiently big such that the overfitting of trees will be countered by the voters' overrule.

Initially, six tests are deployed using the mentioned methods, similar to the tests of [24]. The whole set is split into a train and test set, being two thirds and one third, respectively. These sets are shuffled using seeds 234, 847, 392, 721, 394, 123, respectively, for TEST1-TEST6.

For the second experiment, the feature selection methods are, similar to [24], combined using the union and intersection operations. To combine the features, first, the principal axes are calculated for all feature selection methods. These axes will denote which principal components of the feature selection methods should be combined. To check when two principal axes are equal to each other, the dot product of the unit vectors (principal axes) is taken. This is the cosine similarity, namely being $\cos(\theta) = \frac{A \cdot B}{||A|| * ||B||}$, for two principal axes A and B. The principal axes will be pointing in the same direction if $cos(\theta) = 1$, so to check whether two vectors are similar, we use the condition $1 - cos(\theta) < \epsilon$, for a value of $\epsilon \to 0$.

For the union operation for two feature selection methods, all of the principal components of the first method will be added to a set. After this, we iterate through the principal axes of the second feature selection method. If we encounter a principal axis which is not an axis in the first feature selection method, we add the corresponding principal component to the set. As for the intersection operation for two feature selection methods, we start with an empty set. We iterate through the principal axes of the first method. For every axis, we check whether it is included in the principal axes of the second method - according to the condition above - and add the corresponding principal component to the set.

In the second experiment, the power set of the feature selection methods is calculated, filtering out any subsets in the power set if the cardinality of the subset is smaller than two. Every subset in the power set is then fed into the union and intersection functions. This way, all possible combinations of the feature selection methods will be combined using the union and intersection functions. The resulting features are hereafter fed into the RFR.

The RFR has a variable tree depth, since the selected amount of features can differ significantly based on the used feature selection methods and the union or intersection operators. Let z be the amount of features in a combination, then the depth of the tree will be set to $1\frac{1}{3}z$. If $z = 15$, the depth of the tree will be equal to 20. This makes the second experiment comparable to the first experiment, where a tree depth of 20 was used too.

TABLE I
SAMPLING EXPERIMENT MSE AVERAGES FOR SEQUENTIAL SAMPLING AND PRIORITIZED SAMPLING. ALL VALUES IN THE TABLE ARE MULTIPLIED BY 10^4. THE AMOUNT OF SAMPLES THAT HAVE BEEN GENERATED ARE DENOTED IN PERCENTAGES, WHERE 100% DENOTES 2087 SAMPLES.

	RFR	PCA	IPCA	KPCA	ICA	a	% of entire data set
SSSF[1]	38.37	38.99	39.07	37.46	39.18	-	100
SSMF[2]	39.01	38.63	38.73	37.99	38.95	-	100
PSSF[3]	16.46	16.21	15.96	18.40	15.69	3	86
PSMF[4]	15.36	14.86	14.95	20.36	15.13	2.75	86
PSMF[4]	14.94	14.65	14.73	19.60	14.48	3	86
PSMF[4]	14.59	14.04	13.96	18.76	14.39	4	86
PSMF[4]	15.73	15.43	15.54	20.13	15.54	5	86
PSMF[4]	19.78	19.14	19.07	19.11	19.64	∞	72

[1] Sequential sampling, single feature per day
[2] Sequential sampling, multiple features per day
[3] Prioritized sampling, single feature per day
[4] Prioritized sampling, multiple features per day

Additionally, the same random seeds are used and thus the same amount of tests are deployed per run. There are a total of four runs, where each run uses $\epsilon = 10^{-k}$ for $k = 1, 2, 3, 4$ for runs 1 to 4, respectively. Only the top three performing methods of every run are displayed along with the acquired MSEs for every test.

V. RESULTS

In this section we discuss the results acquired from the experiments mentioned in Section IV. We evaluate the usefulness of the sampling method, the suitability of the time series for this research and discuss the performances of the feature selection methods.

A. Sampling

Calculating the arithmetic mean of the whole data set (Figure 1) equals $-6.7 * 10^{-5}$ whereas the mean of the prioritized sampling simulation equals $4.6 * 10^{-5}$. If we look at the means of the stationary and non-stationary values, we get $2.5 * 10^{-5}$ and $-3.2 * 10^{-4}$, respectively. The mean of the stationary values is higher, which is why the mean of the prioritized sampling simulation is higher too, because we sample more stationary values than non-stationary ones. Since these means differ in value and depend on the time t the measurements are made at, we conclude that $\mu_x(t)$ is dependent on t and therefore, according to Definition 2, the used time series is non-stationary. Additionally the stationary values are tested using the ADF and KPSS tests to check whether our observations were correct. Both tests gave the results of $p = 3.2 * 10^{-28}$ and $p = 0.1$, respectively. This means that we reject the null hypothesis for the ADF test and fail to reject the null hypothesis for the KPSS test. The implication in both cases is that the observed components are indeed stationary. Therefore, this data set is suitable for the research. Additionally, the count of the amount of sampled stationary and non-stationary values are also tracked, in which if we divide the former by the latter, we get the ratio of 2.98, which approximates $a = 3$. The seed used for the simulation was 6437. These results confirm our expectations for the acquired samples.

Furthermore, results of sequential sampling and prioritized sampling have been compared to each other. The sequential samples have been generated by using a for-loop, starting the loop at 2011 and ending it at 2019. Every iteration, the log returns and the trading volumes up to ten days (batch size) ahead are added to the samples, such that instances like $l_t, v_t, ..., l_{t+10}, v_{t+10}$ are acquired, for a log return l_t and trading volume v_t at time t. There have been eight runs in total for the RFR, PCA, IPCA, KPCA and ICA methods, of which the averages are listed in Table I. Every run, six different seeds have been used, which are the same seeds as mentioned in Section IV. Of the eight runs, two runs have been been of sequential sampling and six runs have been of prioritized sampling. For both sampling methods, one run was made where only the log returns are used as the feature per day (SF). For all other runs, both the log returns and the trading volumes are used as features per day (MF). The runs for prioritized sampling with multiple features per day have been run for $a = 2.75$, $a = 3$, $a = 4$, $a = 5$ and $a \to \infty$.

Looking at Table I, it can be seen that the prioritized sampling method using multiple features per day for $a = 4$ had the lowest MSEs out of all entries. It is notable that for $a \to \infty$, the MSEs were relatively large, compared to the other MSEs for prioritized sampling. This, as anticipated in Section IV-A, is the result of biasing our models towards stationary values, while ignoring the

non-stationary ones. The fact that our model contains more stationary values than non-stationary ones, helps the model to predict for regions alike. There is a clear difference in the MSEs obtained from using $a = 2.75$ and $a = 3$, where in the former, the non-stationary values are sampled with the probability of $\frac{1}{3.75}$, which is the actual fraction of the non-stationary values in the entire data set. This indicates that, indeed, by sampling non-stationary values with a lower probability, lower MSEs are acquired. Interestingly enough, the value for $a = 2.75$ should result in an approximation of sequential sampling, since in both cases the non-stationary values get sampled with the same frequency, however, this does not seem to be the case when looking at the MSEs. This could be caused by the way the samples are fed into the feature selection methods: the prioritized samples are ordered randomly, while the sequential samples are ordered chronologically.

Additionally, there seems to be an optimal value for a, which is $a = 4$, looking at the acquired results. A theoretical explanation of this optimal value could be, that for low values of a, the non-stationary values tend to be sampled too frequently and bias the data towards these values. For high values of a, the data gets biased towards the stationary values. In both cases, the biases result in a higher MSE. As of now, there is no algorithm to determine an optimal value for a, which future research may provide more clarity on.

B. Correlation

As mentioned in Section IV-B, the Kolmogorov-Smirnov test is applied to the log returns and the proposed Student-t distribution. The calculated test statistic $D_{N,M}$ is $10.50 * 10^{-3}$ and the corresponding p-value is 0.9. Because of this, $D_{N,M} < 10.77 * 10^{-3}$ and $p > 0.05$, which means that we fail to reject the null hypothesis and that the log returns are most likely distributed as the proposed Student-t distribution.

C. Single feature selection methods

The average MSEs for the prioritized sampling with multiple features per day are displayed for various values of a in Table I. From these results, it can be seen that the PSSF method for $a = 3$ performed better than the PSMF method for $a = 2.75$ and $a > 4$. A very notable point is the improvement provided by prioritized sampling compared to sequential sampling. The lowest MSE is provided by the IPCA method using PSMF for $a = 4$, and compared to the baseline RFR method using SSSF the improvement is

$$\frac{13.96 - 38.37}{38.37} = -63.62\%.$$

It should be, however, noted that all of the feature selection methods use the same amount of features to select, being 15. This has been done for comparability, even though the optimal amount of components for KPCA and ICA may differ. Because the amount of features is the same for every method, the FWMS of the methods will not differ and is hence not included in Table I. Another point worth mentioning is the initial amount of features, which is $9 * 2 = 18$. It may be possible to achieve better results when more features are used.

D. Multiple feature selection methods

In Table II, the results for the second experiment are listed for ϵ values of $\epsilon = 10^{-1}$, $\epsilon = 10^{-2}$, $\epsilon = 10^{-3}$ and $\epsilon = 10^{-4}$. The FWMS of the methods, using $t = 100$, is multiplied by 100. The selection for t is done because we are interested in a significance of two decimals behind the comma, resulting in $t = 10^2$. It is seen that the average of the features generated by the IPCA $\cup$ PCA $\cup$ KPCA method gives the best MSE in all tests, which is, compared to RFR method using SSSF in Table I, being 38.37, an improvement of

$$\frac{13.45 - 38.37}{38.37} = -64.94\%,$$

which is a slight improvement of the improvement mentioned in Section V-C. The MSEs of this combination for $\epsilon = 10^{-1}$ and $\epsilon = 10^{-3}$ are the same, however, the amount of features selected by the union operator is significantly less for the former case, compared to the latter. This can be seen by looking at the FWMS of the two methods, where for $\epsilon = 10^{-1}$, the FWMS is smaller than for $\epsilon = 10^{-3}$. Furthermore, it should be noted that all of the top three combinations listed in Table II, regardless of the used error ratio ϵ, are all feature selection methods under the union operator. It is, however, notable that the top three methods for all tests for different values for ϵ are all the same. A smaller value for ϵ does, according to the results, not necessarily result in a lower average MSE. However, the lowest FWMS score is given by the IPCA $\cup$ KPCA method for $\epsilon = 10^{-1}$, which means that this method yields the lowest MSE for the lowest amount of features, relatively. In particular, the improvement over the baseline provided by this method is 64.92%.

The intersection operator, in contrast to the results achieved by [24], performed worse than the union operator. A sensible explanation as to why their intersection operator yielded the best results is that features will be chosen more selectively, compared to the union operator. This, however, was not the case in our results. One of

TABLE II
TOP THREE LOWEST MSE RESULTS FOR COMBINATIONS OF FEATURE SELECTION METHODS FOR $\epsilon = 10^{-1}$, $\epsilon = 10^{-2}$, $\epsilon = 10^{-3}$ AND $\epsilon = 10^{-4}$. ALL VALUES IN THE TABLE ARE MULTIPLIED BY 10^4.

	$\epsilon = 10^{-1}$			$\epsilon = 10^{-2}$			$\epsilon = 10^{-3}$			$\epsilon = 10^{-4}$		
	IPK[1]	IK[2]	PK[3]	IPK[1]	IK[2]	PK[3]	IPK[1]	IK[2]	PK[3]	IPK[1]	IK[2]	PK[3]
TEST1	13.69	13.58	13.71	13.73	13.77	13.79	13.59	13.77	13.79	13.60	13.77	13.79
TEST2	13.95	14.01	14.21	14.06	14.05	14.16	14.06	14.05	14.16	14.11	14.05	14.16
TEST3	12.28	12.32	12.30	12.20	12.32	12.23	12.21	12.22	12.27	12.21	12.22	12.27
TEST4	13.62	13.65	13.60	13.67	13.62	13.80	13.69	13.62	13.80	13.72	13.62	13.80
TEST5	13.36	13.33	13.64	13.37	13.43	13.63	13.37	13.42	13.63	13.40	13.42	13.63
TEST6	13.76	13.84	13.92	13.78	13.79	13.87	13.79	13.79	13.87	13.84	13.78	13.87
Avg. MSE	13.45	13.46	13.56	13.47	13.50	13.58	13.45	13.48	13.59	13.47	13.48	13.59
Avg. Features	33	28	27	38	30	30	39	30	30	39	30	30
FWMS ($*100$)	67.73	61.90	125.13	90.48	89.23	161.11	80.04	76.94	173.42	92.86	76.94	173.42

[1] IPCA $\cup$ PCA $\cup$ KPCA
[2] IPCA $\cup$ KPCA
[3] PCA $\cup$ KPCA

the causes might be the difference in feature selection methods and another one being the amount of selected features. We have used a total of 15 features for every feature selection method, so an intersection between the methods will only result in less than 15 features. Even though we were selective in our choice of features, the resulting features were too few to yield accurate predictions. This also opens up opportunities for future research, which, as mentioned in Section V-C, may involve the usage of more features than the log returns and the corresponding trading volumes.

The union operator, however, proves to be useful in filling in for the missed explained variance by other feature selection methods. Additionally, we acquire more features and thus provide the RFR with more information to base its predictions on. This, however, might not be sensible right away, because essentially, the principal components captured by the used methods will try to explain the same variance of the data, but each will do so in a different way. It might thus not be obvious that the union operator will actually help with the prediction. An explanation as to why we actually have received better results, may have to do with the RFR, which does not make any assumptions on the underlying distribution of the input data. We may then hypothesize that the principal components generated by the methods increase the accuracy of the prediction, because of the different way they explain the variance and therefore add information to the predictor. These features may even be used by the RFR to add splitting variables for validation to its CARTs.

VI. FUTURE WORK

One of the parameters introduced in this paper is the a parameter, which determines the amount of bias the prioritized sampling simulation has towards the stationary or non-stationary data. This parameter has been tuned manually, however, future research could provide a method to systematically find the optimal value for a. Additionally, the selection of the non-stationary parts has been done through observation. This process could also be optimized, mainly through the field of extreme value theory, which provides methods like the Peaks-Over-Threshold method to determine extreme values, in which case the non-stationary data could be seen as those. Additionally, the amount of samples drawn from the data for $a \neq \infty$ may also be systematically defined in future research.

With the rise of deep learning techniques and deep feature selection methods, newer methods of feature selection and their combinations are introduced. One of these combination methods are discussed by [18], which introduces rankings to features and combines only the top k ranked features of two (or more) methods. Future research may involve the usage of this combination technique. Another point, as mentioned in Section V, is the usage of more features to achieve a better accuracy and to fully make use of the feature selection methods. In this research, the PCA, IPCA, KPCA and ICA methods were used as feature selection methods. Future work may also explore more feature selection methods, apart from PCA, such as a Continuous Restricted Boltzmann Machine (CRBM) or Nonzero Matrix Factorization (NMF). Furthermore, the hypothesis mentioned in Section V-D, may be researched in future work by comparing the results acquired from different predictors using the same input data, being the union of the used feature selection methods.

VII. Conclusion

This research was aimed at the usage of multiple feature selection methods in the short-term prediction of non-stationary time series. The non-stationarity of the time series is partially omitted by the use of prioritized sampling, in which batches from stationary parts of the data are drawn with a higher probability than batches from non-stationary parts. This way, the effects the non-stationary values have on the entire data set are also taken into consideration, but have lesser weight in the prediction. The chosen parameter for sampling is $a = 4$, for which batches from the non-stationary data will be sampled with a probability of $\frac{1}{5}$ and batches from the stationary data will be sampled with a probability of $\frac{4}{5}$. The results of our tests have confirmed our hypothesis to be true, namely, that forecasting non-stationary time series can be done more accurately with the usage of prioritized sampling and feature selection methods. As mentioned by [14], an additional variable has been used in the prediction, being the trading volume in the case of the stock market. This variable has been analyzed on its correlation with the log returns of the sampled data and has been found to be Student-t correlated with the log returns. Finally, multiple feature selection methods are deployed on the data, after which the features are fed into an RFR. The best features were given by the IPCA $\cup$ PCA $\cup$ KPCA method, which gave an improvement of 64.94% compared to the MSE of an RFR for sequential sampling using one feature per day. The lowest FWMS is given by the IPCA $\cup$ KPCA method, which is the model that achieves the lowest MSE using as few features as possible, relatively. Its improvement over the same baseline is 64.92%. All in all, the proposed methods can be used to forecast non-stationary time series, which contain stationary component. Because of the use of feature selection methods and a random forest, more information about the prediction can be acquired and can be acted upon accordingly.

References

[1] Jacob Benesty et al. "Pearson correlation coefficient". In: *Noise reduction in speech processing*. Springer, 2009, pp. 1–4.

[2] Leo Breiman. "Bagging Predictors". In: *Machine Learning* 24.2 (Aug. 1996), pp. 123–140. ISSN: 1573-0565. DOI: 10.1023/A:1018054314350. URL: https://doi.org/10.1023/A:1018054314350.

[3] Leo Breiman. "Random Forests". In: *Machine Learning* 45.1 (Oct. 2001), pp. 5–32. ISSN: 1573-0565. DOI: 10.1023/A:1010933404324. URL: https://doi.org/10.1023/A:1010933404324.

[4] Peter J Brockwell, Richard A Davis, and Matthew V Calder. *Introduction to time series and forecasting*. Vol. 2. Springer, 2002.

[5] Pierre Comon. "Independent component analysis, a new concept?" In: *Signal processing* 36.3 (1994), pp. 287–314.

[6] Rainer Dahlhaus et al. "Fitting time series models to nonstationary processes". In: *The annals of Statistics* 25.1 (1997), pp. 1–37.

[7] William H Greene. "Econometric analysis 4th edition". In: *International edition, New Jersey: Prentice Hall* (2000), pp. 201–215.

[8] Robert S. Hudson and Andros Gregoriou. "Calculating and comparing security returns is harder than you think: A comparison between logarithmic and simple returns". In: *International Review of Financial Analysis* 38 (2015), pp. 151–162. ISSN: 1057-5219. DOI: https://doi.org/10.1016/j.irfa.2014.10.008. URL: http://www.sciencedirect.com/science/article/pii/S1057521914001380.

[9] Aapo Hyvärinen and Erkki Oja. "Independent component analysis: algorithms and applications". In: *Neural networks* 13.4-5 (2000), pp. 411–430.

[10] Ian Jolliffe. *Principal component analysis*. Springer, 2011.

[11] David A Kenny and D Betsy McCoach. "Effect of the number of variables on measures of fit in structural equation modeling". In: *Structural equation modeling* 10.3 (2003), pp. 333–351.

[12] FM Dekking C Kraaikamp and HP Lopuhaä LE Meester. *A Modern Introduction to Probability and Statistics*. 2005.

[13] Denis Kwiatkowski et al. "Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root?" In: *Journal of econometrics* 54.1-3 (1992), pp. 159–178.

[14] Martin Langkvist, Lars Karlsson, and Amy Loutfi. "A review of unsupervised feature learning and deep learning for time-series modeling". In: *Pattern Recognition Letters* 42 (2014), pp. 11–24. ISSN: 0167-8655. DOI: https://doi.org/10.1016/j.patrec.2014.01.008. URL: http://www.sciencedirect.com/science/article/pii/S0167865514000221.

[15] Roger J Lewis. "An introduction to classification and regression tree (CART) analysis". In: *Annual meeting of the society for academic emergency medicine in San Francisco, California*. Vol. 14. 2000.

[16] Vittorio Lippi and Giacomo Ceccarelli. "Incremental Principal Component Analysis Exact

implementation and continuity corrections". In: *arXiv preprint arXiv:1901.07922* (2019).

[17] Frank J Massey Jr. "The Kolmogorov-Smirnov test for goodness of fit". In: *Journal of the American statistical Association* 46.253 (1951), pp. 68–78.

[18] Jouni Pohjalainen, Okko Räsänen, and Serdar Kadioglu. "Feature selection methods and their combinations in high-dimensional classification of speaker likability, intelligibility and personality traits". In: *Computer Speech & Language* 29 (Nov. 2013). DOI: 10.1016/j.csl.2013.11.004.

[19] "Predicting stock index increments by neural networks: The role of trading volume under different horizons". In: *Expert Systems with Applications* 34.4 (2008), pp. 3043–3054. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2007.06.023. URL: http://www.sciencedirect.com/science/article/pii/S0957417407002345.

[20] Bernhard Schölkopf, Alexander J. Smola, and Klaus-Robert Müller. "Advances in Kernel Methods". In: ed. by Bernhard Schölkopf, Christopher J. C. Burges, and Alexander J. Smola. Cambridge, MA, USA: MIT Press, 1999. Chap. Kernel Principal Component Analysis, pp. 327–352. ISBN: 0-262-19416-3. URL: http://dl.acm.org/citation.cfm?id=299094.299113.

[21] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. *Nonlinear component analysis as a kernel eigenvalue problem*. 1996.

[22] Carolin Strobl, James Malley, and Gerhard Tutz. "An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests." In: *Psychological methods* 14.4 (2009), p. 323.

[23] Clifton D Sutton. "Classification and regression trees, bagging, and boosting". In: *Handbook of statistics* 24 (2005), pp. 303–329.

[24] Chih-Fong Tsai and Yu-Chieh Hsiao. "Combining multiple feature selection methods for stock prediction: Union, intersection, and multi-intersection approaches". In: *Decision Support Systems* 50.1 (2010), pp. 258–269. ISSN: 0167-9236. DOI: https://doi.org/10.1016/j.dss.2010.08.028. URL: http://www.sciencedirect.com/science/article/pii/S0167923610001521.

[25] Eleni I Vlahogianni, Matthew G Karlaftis, and John C Golias. "Short-term traffic forecasting: Where we are and where we're going". In: *Transportation Research Part C: Emerging Technologies* 43 (2014), pp. 3–19.

[26] Haitao Zhao, Pong Chi Yuen, and James T Kwok. "A novel incremental principal component analysis and its application for face recognition". In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 36.4 (2006), pp. 873–886.