Physica A 483 (2017) 139–155

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Non-parametric causality detection: An application to social media and financial data

Fani Tsapeli [a,*], Mirco Musolesi [b,c], Peter Tino [a]

[a] School of Computer Science, University of Birmingham, Edgbaston B15 2TT Birmingham, UK
[b] Department of Geography, University College London, Gower Street WC1E 6BT London, UK
[c] The Alan Turing Institute, 96 Euston Road, London NW1 2DB, UK

HIGHLIGHTS

- A causal inference approach for time series based on matching design.
- The method can handle high dimensional data without any assumptions about the model class.
- We use our method to assess the causal impact of social media sentiment on traded assets.

ARTICLE INFO

Article history:
Received 19 November 2016
Received in revised form 6 February 2017
Available online 4 May 2017

Categories and subject descriptors:
G.3

Keywords:
Causality
Social media
Stock market
Sentiment tracking
Time-series

ABSTRACT

According to behavioral finance, stock market returns are influenced by emotional, social and psychological factors. Several recent works support this theory by providing evidence of correlation between stock market prices and collective sentiment indexes measured using social media data. However, a pure correlation analysis is not sufficient to prove that stock market returns are influenced by such emotional factors since both stock market prices and collective sentiment may be driven by a third unmeasured factor. Controlling for factors that could influence the study by applying multivariate regression models is challenging given the complexity of stock market data. False assumptions about the linearity or non-linearity of the model and inaccuracies on model specification may result in misleading conclusions.

In this work, we propose a novel framework for causal inference that does not require any assumption about a particular parametric form of the model expressing statistical relationships among the variables of the study and can effectively control a large number of observed factors. We apply our method in order to estimate the causal impact that information posted in social media may have on stock market returns of four big companies. Our results indicate that social media data not only correlate with stock market returns but also influence them.

1. Introduction

We are living in the era of social media, using tools such as Facebook, Twitter and blogs to communicate with our friends, to share our experiences and to express our opinion and emotions. Recently, mining and analyzing this kind of data has emerged as an area of great interest for both the industrial and academic communities. Several studies have examined the

* Corresponding author.
E-mail addresses: t.tsapeli@cs.bham.ac.uk (F. Tsapeli), m.musolesi@ucl.ac.uk (M. Musolesi), p.tino@cs.bham.ac.uk (P. Tino).

http://dx.doi.org/10.1016/j.physa.2017.04.101
0378-4371/

ability of social media to serve as crowd-sensing platforms. For example, authors in [1] demonstrate that social media can monitor the popularity of products or services and predict their future revenues. Evidence has been found that social media can be used to predict election results [2] or even stock market prices [3].

Most studies so far have focused on using social media data as early indicators of real-world events. But to what extent do opinions expressed through social media actually have a *causal* influence on the examined events? For example, are stock market prices influenced by the opinions and sentiments that are reported in social media, or is it the case that stock market prices and sentiments are driven only by other (e.g. financial) factors? Would the results have been different if we could manipulate social media data? In order to answer such questions a causality study is required.

Some recent studies have examined the ability of social media to influence real-world events by applying randomized control trials. For example, authors in [4] examine the effect of political mobilization messages by using Facebook to deliver such messages to a randomly selected population; the effect of the messages is measured by comparing the real-world voting activity of this group with the voting activity of a control group. Similarly, in [5] authors use randomized trials in order to examine the social influence of aggregated opinions posted in a social news website. Randomized control trials are a reliable technique for conducting causal inference studies [6]. However, their applicability is limited since they require scientists to gather data using experimental procedures and do not allow the exploitation of the large amount of observational data. In many cases, it is not feasible to apply experimental designs or it is considered unethical.

In this work, we study the causal impact of social, psychological and emotional factors on stock market prices of big companies using observational data collected through Twitter. Twitter enables us to capture people sentiments and opinions about traded assets and their reactions on related news and events. Previous works have demonstrated that social media data correlate with stock market prices [3,7–9]. These studies were predominantly based on correlation or Granger causality analysis. Granger causality tests the ability of a time-series to predict values of another one [10]. However, it cannot be used to discover real causality. A positive result on a Granger causality test does not necessarily imply that there is a causal link between the examined time-series since both the examined time-series may be influenced by a third variable (*confounding bias*). Multivariate regression techniques can be applied in order to control for confounding bias. Some studies attempt to improve the accuracy of stock market prediction models by applying multivariate regression [11]. However, the focus of these works is on prediction rather than on causal inference. Applying regression models for causal inference suffers from two main limitations. First, stock market prices can be influenced by a large number of factors such as stock market prices of other companies [12,13], foreign currency exchange rates and commodity prices. Such factors may also influence people sentiments. Consequently, to eliminate any confounding bias one is required to include a large number of predictors in the regression model. Estimation of regression coefficients in a model with a large number of predictors can be challenging. When data dimensionality is comparable to the sample size, noise may dominate the 'true' signal, rendering the study infeasible [14]. Second, inaccuracies in model specification, estimation or selection may result in invalid causal conclusions.

Given the limitations of parametric methods, we propose a novel framework for causal inference in time-series that is based on *matching design* [15,16]. This technique attempts to eliminate confounding bias by creating pairs of *similar* treated and untreated objects, i.e. objects with similar values on baseline characteristics that could influence the causality study. Thus, the effect of an event is estimated by comparing each object exposed to an event with a *similar* object that has not been exposed. Matching design bypasses the limitations of regression-based methods since it does not require specification of a model class. However, it cannot be applied in time-series since it assumes that the objects of the study are realizations of i.i.d. variables. We reformulate the concept of matching design to make it suitable for causal inference on time-series data. In our case the time-series collection includes *treatment* time-series X, *response* time-series Y and a set of time-series $\mathbf{Z}$ which contain characteristics relevant to the study. The units of our study correspond to time-samples; the tth unit is characterized by a treatment value $X(t)$, a response value $Y(t)$ and a set of values representing baseline characteristics $\mathbf{Z}(\mathbf{t})$. We assess the causal impact of a time-series X on Y by comparing different units (i.e. time-samples) on Y after controlling for characteristics captured in $\mathbf{Z}$. As explained in Section 3, our methodology assures that the objects are uncorrelated, which is a weaker version of the independence assumption requirement of the matching design. We apply our framework in order to estimate the causal impact that the sentiment of information posted in social media may have on traded assets. In detail, we estimate a daily sentiment index (*treatment* time-series) based on information posted in Twitter and we assess its impact on daily stock market closing prices (*response* time-series) of four big technological companies after controlling for other factors (set of time-series $\mathbf{Z}$) that may influence the study, such as the performance of other big companies.

In summary, the contribution of this work is twofold:

1. We propose a causal inference framework for time-series that can be applied to high-dimensional data without imposing any restriction on the model class describing the associations among the data. We demonstrate, using synthetic data, that our methodology is more effective on detecting true causality compared to other methods that have been applied so far, for causal inference in time-series.
2. We apply our method in order to quantify the causal impact of emotional and psychological factors, captured by social media, on stock market prices of four technological companies. To the best of our knowledge, this is the first study that measures the causal influence of such factors on finance. It should be noted that, since all the examined companies belong to the technological sector, our findings cannot be directly generalized for any company.

The rest of this paper is organized as follows. In Section 2 we discuss the main methodologies that are used for causal inference. In Section 3 we present the proposed framework. In Section 4 we evaluate our approach on synthetic data,

in conjunction with other methods that have been previously applied for causal inference in time-series. Moreover, we apply our method in order to assess the causal impact of information posted in Twitter on stock market prices of specific companies. In Section 5 we discuss some relevant works which attempt to uncover the relationship between social media data and stock market movement. In Section 6 we summarize the advantages and limitations of the proposed approach and we discuss its computational cost. Finally, Section 7 concludes the paper by summarizing our contributions.

2. Background on causal inference

Causal analysis attempts to understand whether differences on a specific characteristic Y within a population of *units* are caused by a factor X. Y is called *response*, *effect* or *outcome* variable and X *treatment* variable or *cause*. *Units* are the basic objects of the study and they may correspond to humans, animals or any kind of experimental objects.

2.1. Potential outcome framework

The key concept on causation theory is that given a unit u, the value of the corresponding response variable $Y(u)$ can be manipulated by changing the value of the treatment variable $X(u)$ [17,18]. In this paper, we will consider $X(u)$ as a binary treatment variable. Hence, there will be two treatments: x_1 for treated units and x_0 for untreated. We denote by $Y_1(u)$ the value of Y when $X(u) = x_1$ and $Y_0(u)$ when $X(u) = x_0$. In order to test the effect of x_1 on unit u, we need to estimate the quantity $Y_1(u) - Y_0(u)$. The fundamental problem of causal inference is that *it is impossible to observe both* $Y_1(u)$ *and* $Y_0(u)$ *on the same unit* u *and, therefore, it is impossible to measure the real causal effect of* x_1 *on the unit.* [17,18]. Thus, the average effect of a treatment X is estimated by comparing a population of objects that received the treatment x_1 with a population that received the treatment x_0 and evaluating the corresponding average values of the effect variable Y. We denote by Y_1 and Y_0 random variables representing the outcome variable when the treatments x_1 and x_0 are applied, respectively. We also define a (random) variable $DY = Y_1 - Y_0$. Then, the average treatment effect (ATE) of x_1 is estimated as the expected value $E\{DY\}$.

The average treatment effect can be estimated only if the following three assumptions are satisfied:

1. the effect differences are i.i.d. realizations of DY.
2. the observed outcome in one unit is independent from the treatment received by any other unit (*Stable Unit Treatment Value Assumption—SUTVA*).
3. the assignment of units to treatments is independent from the outcome (*ignorability*). Ignorability can be formally expressed as $Y_1 \perp\!\!\!\perp X$, $Y_0 \perp\!\!\!\perp X$. The assumption of ignorability requires that all the units have equal probability to be assigned to a treatment. If this assumption does not hold, the units that received a treatment x_1 may systematically differ from units that did not receive such a treatment. In such a case the average value of the outcome variable of the treated units could be different from that of other units, even if the treatment had not been received at all.

In experimental studies, ignorability can be achieved by randomly assigning units to treatments. However, in observational studies this is not feasible. Instead, the average treatment effect can be estimated by relaxing ignorability to *conditional ignorability*. According to conditional ignorability assumption, the treatment assignment is independent from the outcome, conditional on a set of confounding variables $\mathbf{H}$. Variables $\mathbf{H}$ represent baseline characteristics of the units that are considered relevant for the study (e.g. in a medical study that examines the impact of a drug, baseline characteristics could be the previous health condition of the units (in this case patients), their age etc.). Thus, conditional ignorability is expressed as $Y_1 \perp\!\!\!\perp X|\mathbf{H}$, $Y_0 \perp\!\!\!\perp X|\mathbf{H}$. The variables that must be included in the set $\mathbf{H}$ in order to achieve conditional ignorability are also called *confounding variables*. Conditional ignorability cannot be achieved when one or more confounding variables are unobserved. The main limitation of all non-experimental causality studies is that the possibility that important confounding variables are missing cannot be eliminated. Latent variable models [19] and analysis on the sensitivity of the conclusions on missing confounding variables [20] have been proposed to handle this issue.

There are two main methodologies that are applied in order to achieve conditional ignorability: *regression* and *matching* [15]. Regression expresses the outcome variable Y as a function of the treatment variable X and the set of variables $\mathbf{H}$ [21,22]. Linear models are usually applied. Methods based on regression require scientists to correctly specify a regression model. These methods are affected by the typical problems of parametric approaches to causality detection (i.e. model misspecification, overfitting and poor performance on high-dimensional datasets).

Matching comprises a more flexible methodology for causal inference in observational data since it does not require the specification of a model class [16]. Conditional ignorability is achieved by creating sub-population within which the values of the confounding variables $\mathbf{H}$ are the same or similar. Thus, considering a set G of pairs of treated and untreated (*control*) units (u, v) such that $\mathbf{H}(u) \approx \mathbf{H}(v)$, we can estimate the average treatment effect (ATE) as

$$\widehat{E}\{DY\} = \frac{\sum_{(u,v)\in G} Y(u) - Y(v)}{|G|}, \tag{1}$$

where $|G|$ denotes the size of G. Scientists need to assess the degree of similarity between the matched treated and control units. Similarity relation ($\approx$) can be assessed by estimating the standardized mean difference for each confounding variable

between matched treated and control units, or by applying graphical methods such as quantile–quantile plots, cumulative distribution functions plots, etc. [23–26]. If sufficient balance has not been achieved, the applied matching method needs to be revised.

2.2. Directed acyclic graphs

Pearl [27–29] proposed the use of directed acyclic graphs (DAGs) for representing causal relationships. In causal graphs, nodes represent the variables of the experiment. If **P** a set of the direct predecessors (*parents*) of a node Y, a direct arrow from a node Q (can represent X or one of the variables of set **Z**) to Y will exist only if $Y \not\perp\!\!\!\perp Q|\mathbf{P} \setminus \{Q\}$. A direct arrow from a node Q to a node Y represents a causal relationship between the two variables (i.e. Q causes Y).

Pearl also introduces a graphical criterion for defining a *sufficient set* of variables that need to be controlled in order to achieve conditional ignorability when testing the causal impact of a variable X on a variable Y (*back-door criterion*). According to this rule, a subset **H** of variables is *sufficient* if no element of **H** is a descendant of X and the elements of **H** block all paths from X to Y that end with an arrow to X (*back-door paths*). The intuition behind this criterion is that back-door paths from X to Y represent spurious associations and therefore need to be excluded in order to obtain unbiased estimation of the causal effect of X on Y.

Causal graphs differ from Bayesian graphs since, in a causal graph, an arrow from Q to Y denotes that Y values would change if we could manipulate the values of Q. However, this hypothesis cannot be tested in observational studies. Instead, a graph is created either by utilizing prior knowledge about the structure of the model or by conducting conditional independence tests on the observational data [30–33]. The correctness of the graph can be assessed by fitting the observational data to a system of structural equations derived from the graph (i.e. each variable is regressed on all its direct predecessors) [29].

2.3. Causality on time-series

Causality studies on time-series have been largely based on Granger causality [34]. The Granger causality test examines if past values of one variable are useful in prediction of future values of another variable. In detail, a time-series X Granger causes a time-series Y if modeling Y by regressing it on past values of both Y and X results in reduced residual noise compared to a simple autoregressive model. Granger causality does not test real causality since the conditional ignorability assumption is not satisfied, i.e., the values of both treatment variable X and control variable Y may be driven by a third variable. Moreover, it considers only linear relationships. Granger causality has been extended to handle multivariate cases [35] as well as non-linear cases [36,37]. However, as it was mentioned also at the introduction, inaccuracies on model specification may result in misleading conclusions. In [38] the authors propose an additional model check procedure after fitting a model in order to reduce the amount of false positive causality results. Moreover, in [39] authors propose a time-series causality framework based on graph models. The main advantage of the proposed method is the ability to model latent variables (i.e. unobserved confounding variables). However, this method performs worse than Granger causality for large time-series sample sizes.

Non-parametric approaches (i.e. approaches that do not require the specification of a model class) for causal inference in time-series have also been proposed. Schreiber introduced *transfer entropy* [40] in order to examine whether, given a set of multivariate time-series **S**, the uncertainty about a time-series $Y \in \mathbf{S}$ is reduced by learning the past of a time-series $X \in \mathbf{S}$, when the past of the other time-series in **S** is known. Transfer entropy is a model-free equivalent of Granger causality [41]. The main limitation of this approach is that it requires the estimation of a large number of conditional probability densities which is particularly challenging on continuous datasets [36]. Runge et al. propose the combination of transfer entropy with directed acyclic graphs in order to reduce the number of densities that need to be estimated [42,43]. In detail, causality is estimated by examining whether uncertainty about time-series Y can be reduced by learning the past of X, when the *parents* of Y and X are known. The parents P_Y of a time-series Y are defined as the minimum set of graph-nodes which separate Y with the past of $\mathbf{S} \setminus \{P_Y\}$. Although this modification reduces significantly the number of density estimations that are required, the dimensionality of the dataset may still be high (i.e. the number of parents of Y and X may be very large) which imposes challenges on the estimation of transfer entropy.

3. Proposed mechanism

Given the previously discussed limitations on existing methodologies for causality discovery in time-series, we propose a novel framework that enables causal inference in time-series data that is based on matching design and therefore does not require specification of a model class. The proposed framework also requires only few conditional independence tests, thus it can handle more effectively high-dimensional data. Denote by $Y = \{Y(t_i^y) : i = 0, 1 \ldots, N\}$ and $X = \{X(t_i^x) : i = 0, 1 \ldots, N\}$ the time-series that represent the effect and the cause, respectively and by $\mathbf{Z} = \{\mathbf{Z}(t_i^z) : i = 0, 1 \ldots, N\}$ a set of time-series representing other characteristics relevant for the study. In this study, we consider X as a binary treatment variable. Matching design has been proposed also for non-binary treatments [44,45], however, extending our framework to non-binary cases is out of the scope of this study. Let us also denote by $Y^{(l)}, X^{(l)}$ and $\mathbf{Z}^{(l)}$ the l-lagged versions of the time series Y, X and **Z**, respectively (i.e., if $X(t_i^x)$ the ith sample of X, $X^{(l)}(t_i^x) = X(t_{i-l}^x)$). At Fig. 1 we provide

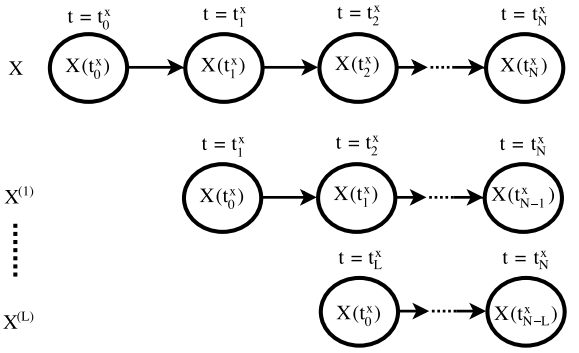


Fig. 1. Graphical representation of time-series X along with its first L lagged versions.

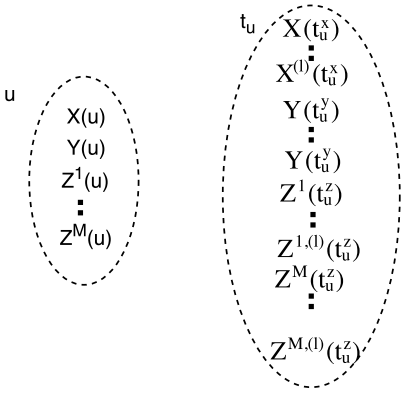


Fig. 2. Graphical representation of units. On the left side, u represents a unit on a traditional causality study, characterized by its treatment value $X(u)$, its response value $Y(u)$ and M other characteristics $Z^1(u), Z^2(u) \ldots, Z^M(u)$. On the right side, t_u represents a unit on our time-series matching design framework. The unit is characterized by the time-series values of set $\mathbf{S}$ at uth time-sample.

a graphical representation of time-series $X, X^{(1)}, \ldots, X^{(L)}$. We define a maximum lag value L and a set of time-series $\mathbf{S} = \{Y, Y^{(1)}, \ldots, Y^{(L)}, X, X^{(1)}, \ldots, X^{(L)}, \mathbf{Z}, \mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(L)}\}$.

As we previously discussed at Section 2, the units of a study traditionally correspond to experimental objects and the variables of the study describe the characteristics of the units as well as the treatment they have received and the corresponding outcome. In our framework, the units of the study correspond to time-samples of the set of times-series $\mathbf{S}$. For example, consider a study that examines the effects of an industry on the pollution level in a region based on weekly measurements. In this case, a unit of the study corresponds to one week and the variables of the study to weekly pollution measurements, industrial wastes and other relevant characteristics. In Fig. 2 we graphically depict the notion of a unit in our time-series matching design framework in comparison with the traditional notion of unit on causality studies. In the rest of this paper, the terms 'unit' and 'time-sample' will be used interchangeably.

In order to build a graph, we examine the dependencies between the variables X, Y and all the other variables of the set $\mathbf{S}$. In order to examine if two time-series X and Y are independent (assuming that the time-series are stationary in the first two moments) we can estimate the Pearson correlation coefficient as follows:

$$r_{xy} = \frac{\sum_{u=0}^{N} (X(t_u^x) - \bar{X})(Y(t_u^y) - \bar{Y})}{\sqrt{\sum_{u=0}^{N} (X(t_u^x) - \bar{X})^2 \sum_{u=0}^{N} (Y(t_u^y) - \bar{Y})^2}}$$

with $\bar{X}$, $\bar{Y}$ the sample means of X, Y respectively. Vanishing correlation could be considered as indication of independence between the examined time-series. Alternatively, Spearman rank correlation or mutual information could be used in order to examine the dependencies between time-series.

In a directed acyclic graph representing a Bayesian network, a arrow from a variable W to a variable Q is added only if Q is dependent of W, conditional on all direct predecessors of Q. In our graph representation, we relax this condition as follows:

A arrow from a lagged node $W^{(l)}$ (including lag 0) to a non-lagged node Q exists if:

- $W^{(l)}$ precedes temporally Q, i.e., $t_u^w < t_u^q$, for any u; and

- $Q \not\perp\!\!\!\perp W^{(l)} | \mathbf{P}^m \cap (W, W^{(1)}, \ldots, W^{(m)})$, where $\mathbf{P^m}$ is the set of the direct predecessors of Q with maximum lag m and $m < l$.

Thus, in our graph representation, a direct edge between two nodes indicates a dependence but not a necessarily a causal link. Causality will be examined by applying the matching design framework, where the direct predecessors of the treatment and outcome time-series will serve as the confounding variables of the study. The main advantage of the proposed framework is the requirement of a significantly lower number of densities estimations and conditional independence testing compared to other causal inference approaches on time-series [40,42,43]. In what follows we will discuss the details of our methodology and how the three general assumptions of causality studies (discussed in Section 2) are addressed.

Data: The set of time-series **S**
Result: The set of confounding variables **H**
`/* Find the parents of Y. */`
P1 := `predecessors(`**S**, Y`);`
`/* Find the parents of X. */`
P2 := `predecessors(`**S**, X`);`
`/* Find the common parents of X, Y. */`
H := **P1** ∩ **P2**;
`/* This procedure returns a set P of the direct predecessors of node Q. P is a subset of S. */`
`Procedure(predecessors(`**S**, Q`))`
P := {};
for *i=0 to L* **do**
  `/* For all zero-lagged time-series */`
  for *all* $S^{(0)} \in$ **S do**
    `/* Find the lagged versions of S which are also parents of Q. */`
    B := $(S^{(0)}, \ldots, S^{(i-1)}) \cap$ **P**;
    if $(Q \not\perp\!\!\!\perp S^{(i)} |$ **B** ***and*** $S^{(i)}$ *precedes* Q **then**
      P := **P** $\cup S^{(i)}$;
    end
  end
end
return **P**;

Algorithm 1: Defining the set of confounding variables.

Conditional ignorability assumption: We apply the Algorithm 1 in order to find the set of time-series **H** that need to be controlled in order to satisfy the conditional ignorability assumption. According to our method, the resulted set contains all the direct predecessors that nodes X and Y have in common. In Fig. 3, we depict the resulted set **H** of an example graph comprised by time-series X, Y and W as well as its lagged versions, with maximum lag $L = 2$. The parents of X and Y are selected by conducting conditional independence tests as described in Algorithm 1. For example, the arrow from $X^{(1)}$ to X denotes that $X \not\perp\!\!\!\perp X^{(1)}$ and the arrow from $X^{(2)}$ to X that $X \not\perp\!\!\!\perp X^{(2)} | X^{(1)}$. Similarly, the arrow from W to X denotes that $X \not\perp\!\!\!\perp W$ and the arrow from $W^{(1)}$ to X denotes that $X \not\perp\!\!\!\perp W^{(1)} | W$. The lack of arrow from $W^{(2)}$ to X denotes that $X \perp\!\!\!\perp W^{(2)} | W, W^{(1)}$. **H** includes all the common parents of X and Y. Thus, all the variables that are correlated both with X and Y time-series are included; hence, the set **H** is sufficient. However **H** may include also redundant time-series, i.e., some of the time-series included at **H** may not correlate with X or Y conditional to a subset of **H**. In causality studies based on regression, including redundant predictors on the model could result in overfitting and would jeopardize the validity of the conclusions. Moreover, the application of methods based on conditional independence tests using information theoretic approaches would be challenged by the inclusion of redundant covariates since it would require conditioning on large sets of variables. In contrast, studies based on matching are less affected by the inclusion of redundant confounding variables (spurious correlations). Several methods that enable matching on a large number of confounding variables have been proposed [46–48]. In addition, scientists are able to apply balance diagnostic tests in order to assess if any confounding bias has been adequately eliminated; consequently, false conclusions due to confounding bias can be diminished. Following the matching design, the set of time-series **H** is controlled by creating a set of pairs of time-samples G where each uth time-sample with a positive treatment value $X(t_u^x)$ is matched with a v^{th} time-sample with zero treatment $X(t_v^x)$ such that $\mathbf{H}(t_u^h) \approx \mathbf{H}(t_v^h)$.

Stable unit treatment value assumption: Denote by **P** the set of time-series that are direct predecessors of the effect variable Y. Assuming $X \in \mathbf{P}$ (if not, X is independent of Y and therefore there is no causation), the assumption is violated if $X^{(l)} \in \mathbf{P}$ and $X^{(l)} \notin \mathbf{H}$, for $l > 0$. Since units correspond to time-samples, $X^{(l)} \in \mathbf{P}$ implies that the outcome value $Y(t_u^y)$ at time t_u^y depends on the value of the treatment time-series X at time t_{u-l}^x. In order to satisfy the assumption, we modify the **H** set as

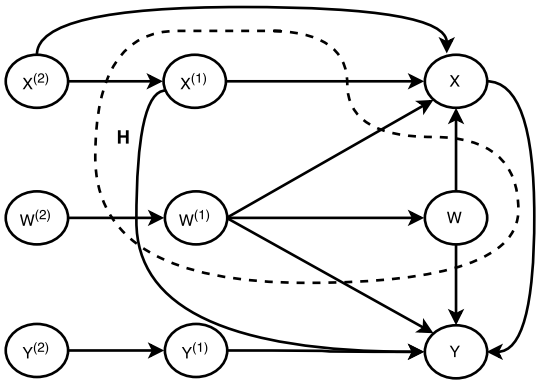


Fig. 3. Example graph depicting the resulted set **H** when the impact of X on Y is examined. At this example, X precedes temporally Y and W precedes X. The maximum examined time-lag L is 2.

follows:

$$\mathbf{H} \coloneqq ((X^{(1)}, \ldots, X^{(L)}) \cap \mathbf{P}) \cup \mathbf{H}, \tag{2}$$

satisfying $Y(t_u^y) \perp\!\!\!\perp X(t_v^x)|\mathbf{H}(t_u^h), \forall u \neq v$.

i.i.d. Assumption: Denote by Y_1 the value of the outcome variable for the time-samples that have a positive treatment value and with Y_0 for time-samples with zero treatment value. The average causal effect is estimated as $\widehat{E}\{Y_1 - Y_0|\mathbf{H}\}$. In order to enable statistical inference, the variable $\Delta Y \coloneqq Y_1 - Y_0|\mathbf{H}$ needs to be i.i.d. If $\mathbf{P}$ the set of direct predecessors of Y, the outcome value $Y(t_u^y)$ of each time-sample t_u^y will depend on the outcome value $Y(t_{u-l}^y)$ if there is a time-series $Y^{(l)} \in \mathbf{P}$. In case that $Y^{(l)} \notin \mathbf{H}$, the i.i.d. assumption would be violated. In order to satisfy this assumption, we modify the set of time-series $\mathbf{H}$ as follows:

$$\mathbf{H} \coloneqq ((Y^{(1)}, \ldots, Y^{(L)}) \cap \mathbf{P}) \cup \mathbf{H}. \tag{3}$$

Causal inference will be performed by matching on the modified set of time-series $\mathbf{H}$ thus, the variable $\Delta Y \coloneqq Y_1 - Y_0|\mathbf{H}$ will be i.i.d.

4. Evaluation

4.1. Evaluation with synthetic data

In order to demonstrate the potential of our approach we assess its effectiveness in detecting causal relationships on linear and non-linear synthetic data. We also compare our approach with a multivariate Granger causality model and with an information theoretic approach based on Runge's framework [43] and we demonstrate that our method is more efficient on avoiding false causal conclusions. We denote with $X = \{X(t_u^x) : u = 1, 2, \ldots, N\}$ and $Y = \{Y(t_u^y) : u = 1, 2, \ldots, N\}$ the treatment and outcome time-series respectively and with $\mathbf{Z} = \{\mathbf{Z}(t_u^z) : u = 1, 2, \ldots, N\}$ a set of M confounding variables. We also assume that $t_u^z < t_u^x < t_u^y, \forall u$. The relationships among X, Y and $\mathbf{Z}$ are described by the following model:

$$X(t_u^x) = h_{xx}(X(t_{u-1}^x)) + f_{xz}(\mathbf{Z}(t_u^z)) + \epsilon_x(t_u^x) \tag{4}$$

$$Y(t_u^y) = h_{yy}(Y(t_{u-1}^y)) + f_{yz}(\mathbf{Z}(t_u^z)) + f_{yx}(X(t_u^x)) + \epsilon_y(t_u^y) \tag{5}$$

$$Z^i(t_u^z) = h_{z^i}(Z^i(t_{u-1}^z)) + \epsilon_{z^i}(t_u^z), \quad \forall Z^i \in \mathbf{Z}, \tag{6}$$

where $\epsilon_x(t_u^x)$, $\epsilon_y(t_u^y)$ and $\epsilon_{z^i}(t_u^z)$ are i.i.d. Gaussian noise variables with zero mean and std. dev. equal to $20 + 2 \cdot M$, $10 + 2 \cdot M$ and 10, respectively.

We consider the following four cases:

Case 1. The model is linear. Thus, $f_{xz}(\mathbf{Z}(t_u^z)) = \sum_i \alpha_{xz,i} \cdot Z^i(t_u^z)$, $h_{xx}(X(t_{u-1}^x)) = \alpha_{xx} \cdot X(t_{u-1}^x)$ $h_{yy}(Y(t_{u-1}^y)) = \alpha_{yy} \cdot Y(t_{u-1}^y)$, $f_{yz}(\mathbf{Z}(t_u^z)) = \sum_i \alpha_{yz,i} \cdot Z^i(t_u^z)$, $f_{yx}(X(t_u^x)) = \alpha_{yx} \cdot X(t_u^x)$, $h_{z^i}(Z^i(t_{u-1}^z)) = \alpha_{z^i} \cdot Z(t_{u-1}^z)$.

Case 2. We apply the linear model of Case 1, but we set $f_{yx}(X(t_u^x)) = 0$. In this case the treatment time-series X does not have any causal impact on the outcome time-series.

Case 3. The dependences between the confounding variables and the treatment and effect variables are non-linear. In particular,

$$f_{xz}(\mathbf{Z}(t_u^z)) = \sum_i \alpha_{xz,i} \cdot (Z^i(t_u^z))^2$$

$$f_{yz}(\mathbf{Z}(t_u^z)) = \sum_i \alpha_{yz,i} \cdot (Z^i(t_u^z))^2.$$

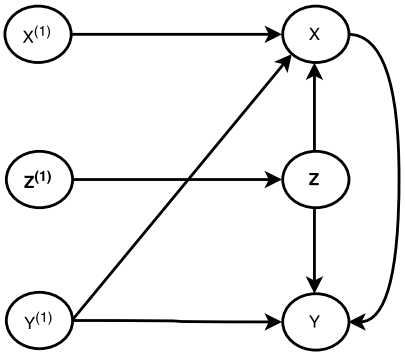


Fig. 4. Resulting graph after applying Algorithm 1 on the synthetic data when $M = 1$ (i.e. there is only one confounding variable). The graph depicts the direct predecessors of nodes X and Y. The set of nodes $\mathbf{H}$ will contain the direct predecessors that nodes X and Y have in common. In the four examined cases X correlates with Y, though in Case 2 and Case 4, this is a spurious correlation due to the set of confounding variables $\mathbf{Z}$. There is also a spurious correlation of node X with node $Y^{(1)}$. X and Y are independent to $\mathbf{Z}^{(1)}$ conditional to $\mathbf{Z}$ and Y is independent to $X^{(1)}$ conditional to X.

We use the linear equations of Case 1 for the rest of the functions.

Case 4. We use the non-linear model of Case 3, but we set $f_{yx}(X(t_u^x)) = 0$. In this case, the multivariate linear Granger causality approach may return positive causality result, even though the treatment time-series $X(t)$ does not have any causal impact on the outcome time-series.

A unit (i.e. time-sample) u of the study is described by the set of time-series values: $S(t_u) := (X(t_u^x), Y(t_u^y), \mathbf{Z}(t_u^z), X^{(1)}(t_u^x), Y^{(1)}(t_u^y), \mathbf{Z}^{(1)}(t_u^z))$. We apply the following three methodologies on the synthetic data generated using the models above in order to assess the causal impact of variable X on Y:

Multivariate Granger Causality (MGC). We apply stepwise regression in order to fit our data to the following model:

$$Y(t_u^y) = a_1 \cdot Y(t_{u-1}^y) + \sum_{l=0}^{(1)} b_l \cdot X(t_{u-l}^x) + \sum_{l=0}^{(1)} \mathbf{c}_l \cdot \mathbf{Z}(t_{u-l}^z) + \delta + \epsilon(t_u^y). \tag{7}$$

We conclude that X causes Y if X or any lagged version of X is included in the regression model.

Conditional Mutual Information Tests (CMI). Following Runge's approach [43] a causal graph is created by performing conditional independence tests using conditional mutual information as described at Section 2.3.

Matching Design for Time-series (MDT). Following the proposed approach, we apply Algorithm 1 in order to find the set of variables $\mathbf{H}$ that needs to be controlled in order to achieve conditional ignorability. The resulted graph is depicted at Fig. 4. $\mathbf{H}$ includes any $Z^i \in \mathbf{Z}$ that correlates both with X and Y. Moreover, we satisfy the i.i.d. assumption by including in $\mathbf{H}$ the time-series $Y^{(1)}$. In order to create groups of treated and untreated units we first transform the time series X into a binary stream $\tilde{X} : \tilde{X}(t_u^x) = 0$, if $X(t_u^x) < \mu_X; \tilde{X}(t_u^x) = 1$, otherwise, where μ_X is the mean of X (i.e. the uth time-sample corresponds to a treated unit if $X(t_u^x) > \mu_X$). Then, we create pairs of treated and untreated units (i.e. time-samples) by applying Genetic Matching algorithm [46]. Genetic matching is a multivariate matching method which applies an evolutionary search algorithm in order to find optimal matches which minimize a loss function. We use as a loss function the average standardized mean difference between the treated and control units for all the confounding variables $H^i \in \mathbf{H}$ which is defined as follows:

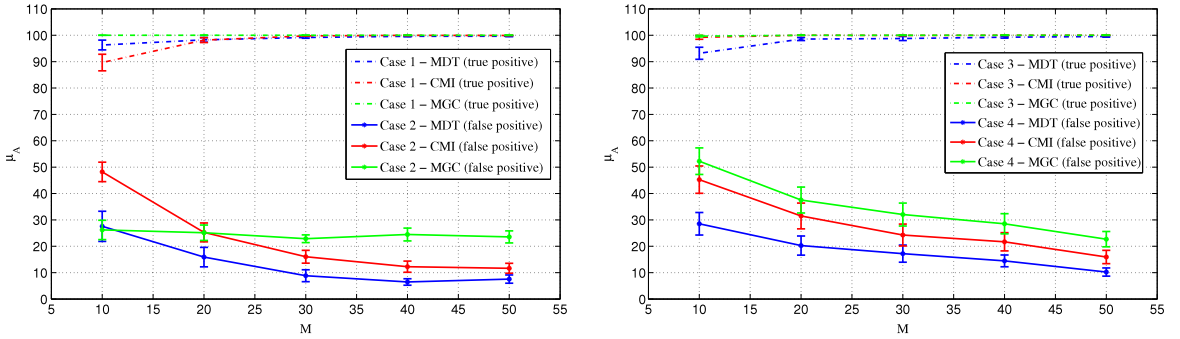
$$SMD_H = \sum_{H^i \in \mathbf{H}} \frac{\sum_{(t_u^h, t_v^h) \in G} |H^i(t_u^h) - H^i(t_v^h)|}{|G| \cdot \sigma_{H^i}} / |\mathbf{H}|. \tag{8}$$

Finally, the average treatment is estimated using Eq. (1) and a t-test is used to examine whether the observed ATE is statistically significant from 0.

We generate 100 samples for each time-series. We vary the number of confounding variables M that are included at set $\mathbf{Z}$ from 10 to 50. In detail, we evaluate the three methodologies for $M = \{10, 20, 30, 40, 50\}$. For each M value, we repeat our study for 30 randomly selected sets of model coefficients (αs). All model coefficients are randomly generated from uniform distribution on $[-4, 4]$ for the linear cases and on $[-1, 1]$ for the non-linear cases. Finally, for each one of the 30 sets of model coefficients we repeat each study for 100 different noise realizations. For the nth noise realization, we define:

$$S_n = \begin{cases} 1 & \text{if } X \text{ was detected as cause of } Y \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

For the kth set of model coefficients we also define $A_k = \sum_{n=1}^{100} S_n$. In Case 1 and Case 3, A_k denotes the number of times that a causal relationship from X to Y is successfully inferred (*true positive*) for the kth set of model coefficients and different noise realizations, while in Case 2 and 4 it denotes the number of times that a causal relationship is falsely inferred (*false*



(a) Linear case. (b) Non-linear case.

Fig. 5. Comparison of the MDT, CMI and MGC causality detection methods on synthetic data.

positive). In Fig. 5 we present the mean value of A_k, μ_A along with the standard error of the mean. According to our results, the proposed causal inference technique reduces significantly the number of false positive causality conclusions while it is slightly less successful on detecting real causality for $M = 10$. Multivariate Granger causality achieves almost 100% accuracy on true causality detection both for the linear (Case 1) and non-linear (Case 3) cases. However, it performs poorly in terms of avoiding false positive conclusions. The performance of all the examined methods improves for larger M values (apart from multivariate Granger causality on the linear cases). This is due to the fact that, by adding more variables on the set **Z**, the dependence of Y and X with each individual $Z^i \in \mathbf{Z}$ is weaker; consequently, although M covariates are used to generate X and Y time-series, for large M values, only a subset of them has significant effect on them. Thus, canceling out the effect of **Z** is easier.

4.2. Causal effect of social media on stock markets

4.2.1. Dataset description

We apply our method in order to investigate whether information about specific companies and people reactions influence stock market prices. We gather this information by analyzing relevant tweets. Twitter enables us to capture people opinions about the target companies, their optimism/pessimism about stock market movements and their reaction to news such as quarterly results announcements or new product launches. Thus, factors related to the company performance and people trust on the company are reflected on Twitter data. Our study considers the daily closing prices of four big tech companies traded on NASDAQ market: Apple Inc., Microsoft, Amazon.com and Yahoo!. We estimate a daily sentiment index for each of these companies by analyzing the sentiment of related tweets. Tweets are gathered using the names and the ticker symbols of the examined companies. In detail, we use the hashtags #apple and #AAPL for Apple Inc., #microsoft and #MSFT for Microsoft, #amazon and #AMZN for Amazon.com and #yahoo and #YHOO for Yahoo![1]. Our study is based on data gathered for four years, from January 2011 to December 2014. We examine whether the sentiment of tweets that are posted before stock market closing time influences the closing prices of the target stocks. In order to eliminate any confounding bias we need to control for factors that may affect both humans sentiment and the target stock prices. Potential influential factor on stocks daily closing prices are their opening prices and their performance during the previous days. Several works have also demonstrated that the performance of other big companies (either local or overseas companies) could influence some stocks (see for example [12,13]). Foreign currency exchange rates may also cause money flows to overseas markets and consequently influence stocks prices. Finally, commodities prices could affect the earnings of companies and, therefore, their stocks prices. More specifically, our study involves the following time-series:

The *response* time-series Y. The difference on the closing prices of the target stocks between two consecutive days. The uth time-sample t_u^y of the time-series Y corresponds to the closing value of the uth day minus the closing value of the previous day.

The *treatment* time-series X. A daily sentiment index that is estimated using tweets related to the target stocks that are posted up to 24 h before the closing time of the corresponding stock market. In order to assure that the values of the treatment variable are driven by information that has been available before the closing time of the target stocks, we omit from the study tweets posted up to one hour before the closing time. Thus, the sentiment index of day u is estimated using all the tweets posted from 4:00 p.m. (ET) time (i.e. the NASDAQ closing time) of day $u - 1$ to 3:00 p.m. (ET) time of day u. Consequently, our treatment variable captures the people sentiment and reactions to news realizes at any time during the day, up to one hour before the stock market closing time. Tweets are filtered using the name of the company and the stock symbol as keywords.

[1] Using the ticker symbol as hashtag for downloading relevant tweets could be problematic for companies with one letter symbol. In such cases, researchers have to identify first the hashtags or names used to identify these companies and search them instead.

Table 1
Accuracy of the text classification for each classification category.

	$P(V_i = 0)$	$P(V_i = 1)$	$P(V_i = 2)$
$i =$ positive	0.05	0.27	0.68
$i =$ neutral	0.03	0.91	0.06
$i =$ negative	0.65	0.29	0.06

The set of time-series Z. We consider the following time-series which might influence our case study:

1. **The difference between the opening and closing prices of two consecutive days**. This time-series is an indicator of the activity of the target stocks at the start of the trading day.
2. **The stock market prices of several major companies around the world**. In our study we include all the components of the most important stock market indexes such as NASDAQ-100, Dow-30, Nikkei 225, DAX and FTSE. The study could be influenced only by factors that precede temporally both the treatment and effect variables. Thus, we use the difference between the opening and closing prices of two consecutive days for stocks that are traded in the USA exchange markets. The closing time of companies traded at the overseas markets precedes the closing time of the USA stock exchange market, thus the time-series for all the overseas companies stocks correspond to the difference on the closing prices between two consecutive days. Although the values of the treatment variable are driven by tweets that are posted both before and after the corresponding values of the time-series that we use to describe the performance of big companies, for convenience, we consider that the uth time-sample t_u^x of the treatment time-series occurs one hour before the USA stock exchange market closing time at day u. Thus, the time-sample t_u^z of any of the time-series that are used to describe the performance of either a USA-based company or an overseas company temporally precedes the uth sample of the treatment time-series.
3. **The daily opening values of foreign currency exchange rates minus the previous day opening values**. We include the exchange rates between dollar and British pound, Euro, Australian dollar, Japanese Yen, Swiss Franc and Chinese Yen.
4. **The difference between the opening values of commodities for consecutive days**. We include the following commodities: gold, silver, copper, gas and oil.

4.2.2. Daily sentiment index estimation

We classify each tweet as negative, neutral or positive using the SentiStrength classifier [49]. SentiStrength estimates the sentiment of a sentence using a list of terms where each term is assigned a weight indicating its positivity or negativity. We updated the list of terms in order to include terms that are commonly used in finance.[2]

Sentiment extraction from text may be inaccurate. Although this issue has been disregarded in previous works [3,8,9], here, in order to account for such inaccuracies on sentiment classification, we estimate a probability distribution function of the daily sentiment instead of a single metric. Let us define a set of three objects $S = \{positive, neutral, negative\}$. Each object $i \in S$ denotes a classification category. Let us also define a random variable V_i as follows:

$$V_i = \begin{cases} 0 & \text{if a negative tweet is classified in class } i \\ 1 & \text{if a neutral tweet is classified in class } i \\ 2 & \text{if a positive tweet is classified in class } i. \end{cases} \tag{10}$$

We derive the probability distribution functions of each random variable V_i, with $i \in S$, based on the classification performance results. We evaluate the performance of the classifier by manually classifying 1200 randomly selected tweets (200 tweets for each one of the four examined companies). The probability distribution functions are presented in Table 1.

Let us define with N_i the number of tweets posted within a day that are classified in category i. We define a random variable $\mathcal{V}_{t_u}$ which corresponds to the sentiment of the uth day as follows:

$$\mathcal{V}_{t_u} = \sum_{i \in S} N_i \cdot V_i. \tag{11}$$

Since 2 is the maximum value of V_i, $\mathcal{V}_{t_u} \in \{0, 1, \ldots, 2 \cdot \sum_i N_i\}$. We estimate the probability distribution of $\mathcal{V}_{t_u}$ by deriving the probability-generating function under the assumption that the real sentiment of a tweet is independent to the sentiment of any other tweet conditional to the observed classification of the tweet sentiment (i.e. the inferred sentiment by SentiStrength). Although the sentiment of a tweet may depend on previously posted tweets, given that the probability of correctly inferring the sentiment of a tweet is independent to the sentiment inference of any other tweet, our assumption is realistic. The probability-generating function of $\mathcal{V}_{t_u}$ is expressed as follows:

$$G_{\mathcal{V}_{t_u}} = \prod_{i \in S} (G_{V_i}(z))^{N_i} = \prod_{i \in S} \left(\sum_{x=0}^{2} p(V_i = x) \cdot z^x \right)^{N_i}. \tag{12}$$

[2] The words list along with the updated words weights can be found here: https://www.cs.bham.ac.uk/tkt357/Publications.html.

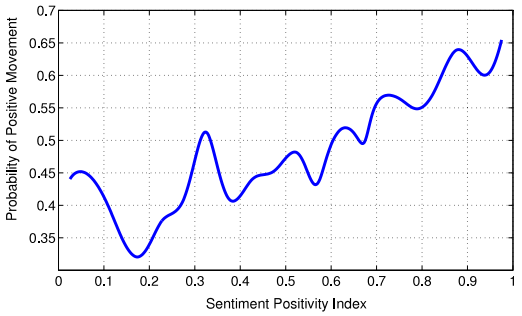


Fig. 6. Probability distribution function of having a positive movement on the traded assets prices conditional to the sentiment of the tweets.

Table 2
Correlation of Y with X, $X^{(1)}$ and $Y^{(1)}$.

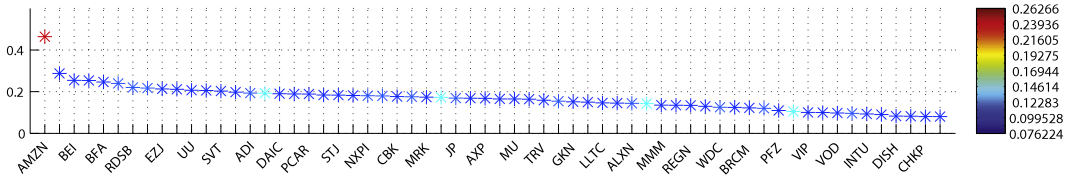
	AAPL	MSFT	AMZN	YHOO
X	0.393	0.155	0.237	0.273
$X^{(1)}$	0.032	0.036	0.012	0.046
$Y^{(1)}$	0.009	−0.003	−0.037	0.031

The probability distribution function of $\mathcal{V}_{t_u}$ is estimated by taking the derivatives of $G_{\mathcal{V}_{t_u}}$. If $\mathcal{N}_{t_u}$ the number of tweets posted the uth day, then, $\mathcal{V}_{t_u} \in \{0, 1, \ldots, \mathcal{N}_{t_u} \cdot M\}$ and the probability that the general sentiment of the uth day is positive is given by the probability $P_{pos}(t_u) = P(\mathcal{V}_{t_u} > \frac{\mathcal{N}_{t_u} \cdot M}{2})$.

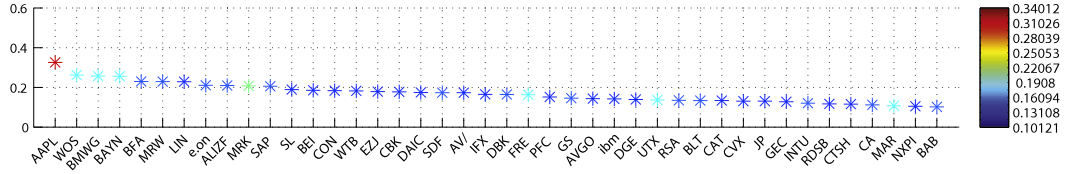
4.2.3. Results

We create a binary treatment variable X by applying thresholds on $P_{pos}(t_u)$. More specifically, a unit that describes the uth day of the study is considered to be treated (i.e. $X(t_u^x) = 1$) if $P_{pos}(t_u^x) \geq P^1_{thresh}$ and untreated (i.e. $X(t_u^x) = 0$) if $P_{pos}(t_u) < P^0_{thresh}$. We conduct our study for three different pairs of thresholds. In detail, we consider a pair of thresholds $T1$, where thresholds P^1_{thresh} and P^0_{thresh} are set to the 50th percentile of X, a pair of thresholds $T2$ where P^1_{thresh} is set to the 60th percentile of X and P^0_{thresh} to the 40th percentile of X and finally a pair $T3$ where P^1_{thresh} and P^0_{thresh} are set to the 70th and 30th percentiles respectively. By increasing the value of P^1_{thresh} and decreasing the value of P^0_{thresh} we eliminate from our study days in which the estimated tweets polarity is uncertain either due to measurement error or because the overall sentiment that is expressed during these days is considered to be neutral. Although discretization of a continuous variable results in information loss which may jeopardize, in some cases, the reliability of the causal inference, we enhance the validity of our conclusions by considering different threshold values.

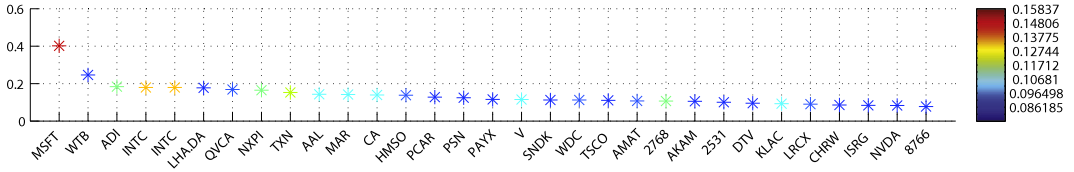
We include in our study all the previously mentioned variables. We found that there is no autocorrelation in our time-series, thus, since there is no dependence of our time-series on their past values, we set the maximum lag L equal to 1. For each of the four target stocks, we applied Algorithm 1 in order to find the set of time-series $\mathbf{H}$ that needs to be controlled. We consider a correlation to be statistically significant if the corresponding p-value is smaller than 0.05. We used Spearman's rank correlation in order to capture potentially non-linear relationships among the examined variables. We found that stock movements are significantly correlated with the sentiment of tweets posted within the same day. Our findings are in agreement with results of other studies [3,7,8]. We also found that stock prices are independent to past tweets sentiment conditional to more recent tweets. This indicates that any effect of tweets on stock prices is instant rather than long-term. Finally, according to our results, the daily movement of the traded assets for the target companies does not correlate with past days movements. This finding is consistent with the weak-form efficient market hypothesis according to which, it is not feasible to predict stock market movements by applying technical analysis. In Table 2 we present the correlation coefficient of the effect variable Y with the treatment variable X and the 1-lagged variables $X^{(1)}$ and $Y^{(1)}$ for each one of the four examined companies. In Fig. 6 we present the empirical probability distribution function of having a positive movement on the traded assets prices conditional to the sentiment of the tweets $P(Y(t_u^y) > 0|X(t_u^x))$. The probability distribution function is estimated using data collectively for the four examined companies. Our results indicate that the probability of having a positive movement on the stock market does not increase linearly with the daily tweets positivity index. Stock market movement is quite uncertain when the positivity index of the tweets ranges between 0.35 and 0.65, while the probability of having a positive movement is increasing for positivity index larger than 0.65. Moreover, we notice a relatively high probability of having a positive movement in days with sentiment positivity index lower than 0.1. Considering that daily tweets sentiment capture the current and past stock market trends, this could be attributed to the fact that investors may consider that it is a good time to invest money when assets prices are low; consequently, this could give lead to an increase of stock market prices.



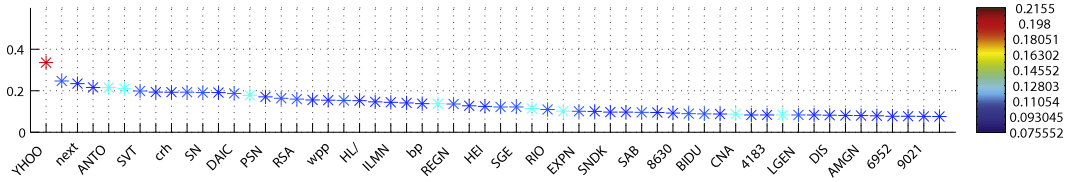
(a) Amazon.



(b) Apple.



(c) Microsoft.



(d) Yahoo.

Fig. 7. Correlation between the confounding variables and the treatment and effect time-series.

Table 3
Number of variables that are included in the set **H** for each of the four examined companies.

	AAPL	MSFT	AMZN	YHOO
Nasdaq-100 comp.	6	21	33	7
Nikkei comp.	1	3	1	13
DAX comp.	18	2	7	10
FTSE comp.	10	3	12	26
Dow-30 comp.	7	3	9	2
FOREX	0	0	0	0
Commodities	0	0	0	0

Moreover, we find that both the effect and the treatment variables correlate with the most recent stock prices of several local and overseas companies. The daily movements of the target stocks correlate with US dollar exchange rates; however, currency exchange rates do not have any impact on the treatment variable. In Table 3 we present the number of variables from each category that will be included in the set **H** for the four target companies and in Fig. 7, we present the correlation coefficients of the treatment and effect time-series with all variables in set **H**. For all the examined stocks, the strongest confounder is their opening prices.

In order to eliminate the effect of the confounding variables we need to match treated and control units with similar values on their set of confounding variables. We create optimal pairs of treated and untreated units by applying Genetic Matching algorithm [46], using as loss function the average standardized mean difference between the treated and control units, as described in Eq. (8), for all the confounding variables. We check if sufficient balance between treated and untreated subjects has been achieved by checking the standardized mean difference for each confounding variable. The remaining bias from a confounding variable is considered to be insignificant if the standardized mean difference is smaller than 0.1 [23,24].

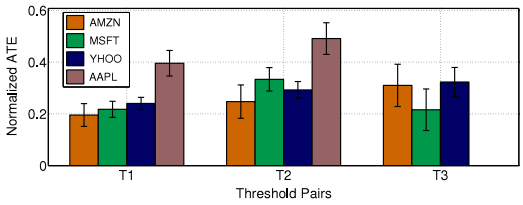


Fig. 8. Normalized ATE for the three threshold pairs.

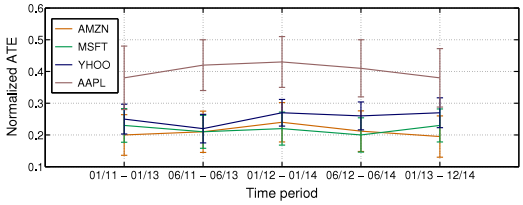


Fig. 9. Normalized ATE for the threshold pair T1. Analysis is conducted in two-year sub-periods using sliding windows with 6 months step.

Table 4
Resulted p-values for the methods MGC and CMI under the null hypothesis that the influence of P_{pos} on Y is zero.

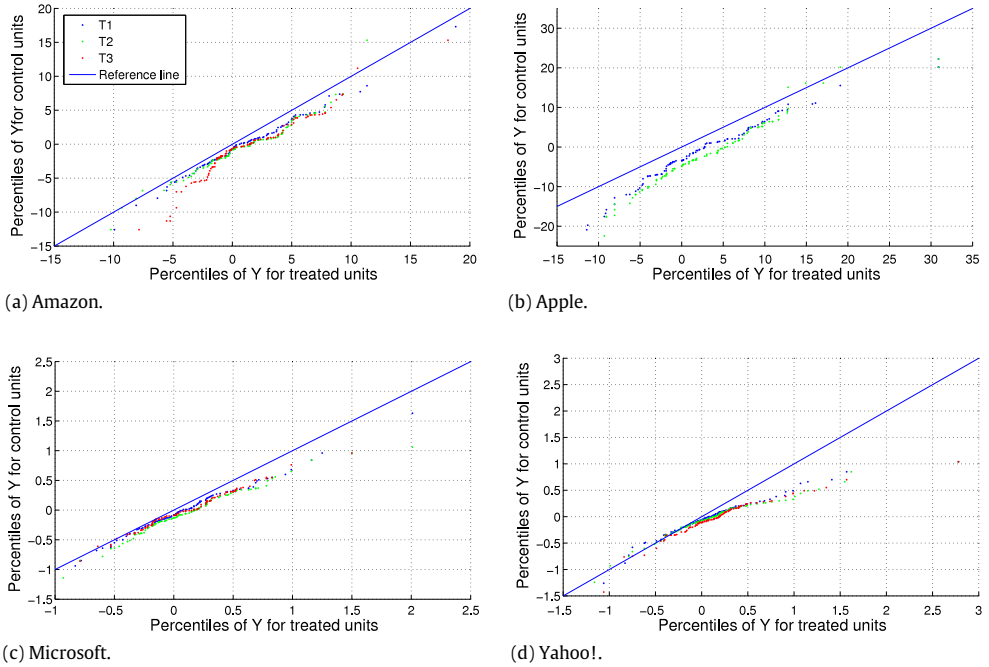
	AAPL	MSFT	AMZN	YHOO
MGC	0.0000	0.0008	0.0000	0.0011
CMI	0.0002	0.0021	0.0001	0.0004

We examine the causal effect of the sentiment of tweets on the target stocks for the three pairs of thresholds. We apply Eq. (1) in order to estimate the average treatment effect (ATE). Under the assumption that the examined treatment has no impact on the effect variable, the ATE would be equal to 0. We use a t-test to assess how significant is the difference of the observed ATE value from 0. In Fig. 8, we present the average treatment effect normalized by the variance of the effect variable Y along with the 95% confidence interval values. According to our results, the effect of the tweets sentiment on the stocks prices of all the examined stocks is statistically significant. We also observe that the causal impact is stronger for larger values of the P^1_{thresh} and smaller P^0_{thresh} threshold values, i.e. the observed difference on the effect variable between the treatment and control groups is larger when we consider only days for which there is less uncertainty on the estimated tweets polarity. For Apple, it was not possible to create balanced treated and control groups for the thresholds pair $T3$. This is due to the fact that the opening prices of the AAPL stocks are very strongly correlated with both the effect and treatment variables and, therefore, there were not enough treatment and control units with similar values on their confounding variables. Since any causal conclusions are not reliable when the treated and control groups are not balanced, we do not present results for Apple for this pair of threshold values. In addition, we repeat our study for different time-periods using a two-year sliding window with six-month step. In Fig. 9 we present our findings for the four examined companies and the first pair of thresholds. According to our results, the difference on the estimated ATE is insignificant for the examined sub-periods. Finally, in Fig. 10 we compare the distributions of the effect variable Y for the treated and control units by plotting their percentiles against each other. Under the hypothesis that the treatment variable has no effect on variable Y, the plot should follow approximately the line $y = x$. However, most of the points of the plot lie below the reference line $y = x$, indicating that the majority of the percentiles of variable Y for the treated units are larger than the corresponding percentiles for the control units.

Finally, we examine the impact of time-series P_{pos} on Y by applying multivariate Granger causality (MGC) and conditional mutual information tests (CMI) as described at Section 4.1. In Table 4 we present the p-values for the two methods under the null hypothesis that the effect of P_{pos} on Y is zero. According to our results, both methods reject the null hypothesis with p-value smaller than 0.05 and confirm our findings that there is a causal link between social media and traded assets prices for the four examined companies. Thus, in this case study, all the examined causal inference methods result in the same conclusion.

4.2.4. Sensitivity analysis

The main limitation of all non-experimental causality studies is that they are based on the assumption that all confounding variables are known. However, in real scenarios there may be unmeasured factors which influence the assignment of units to treatments. In such cases, the conditional ignorability assumption is violated and consequently, any causal inference result may be biased. In our study, we include a large number of potentially influential factors such as the performance of other companies traded assets, commodities prices and currency exchange rates. However, there



(a) Amazon. (b) Apple.

(c) Microsoft. (d) Yahoo!.

Fig. 10. Percentiles of treated units versus percentiles of matched control units.

are other factors, such as inflation rates, political changes or economic policy changes which could influence both people sentiment, captured through Twitter, and traded assets prices. Although such factors may be reflected on the observed confounding variables (e.g. macroeconomic factors such as inflation rates would also affect the prices of other traded assets and consequently, the observed Twitter sentiment may be independent to inflation rates conditional to the performance of other assets included in the study), there may still be some bias due to unobserved factors.

A sensitivity analysis can be conducted in order to assess how the results of the study would be influenced in the presence of unmeasured confounding variables [20,50]. In detail, let us denote with π_{t_u} the probability that a unit t_u corresponding to the uth day is assigned to a treatment (i.e. $X(t_u^x) = 1$) and $O_{t_u} = \pi_{t_u}/(1 - \pi_{t_u})$ the odds of the unit to receive a treatment. Then, we denote with $\Gamma = O_{t_u}/O_{t_v}$ the ratio of the odds of two units t_u, t_v. If $\Gamma = n$, the unit t_u is n times more likely to receive a treatment than unit t_v due to unobserved factors. Under the conditional ignorability assumption (i.e. units are equally likely to receive a treatment conditional to their observed characteristics), the ratio Γ should be equal to 1 for two matched time-samples t_u and t_v.

In [50], Rosenbaum applies the Wilcoxon's signed rank test [51] for the resulted matched treated and control pairs of a causality study under the null hypothesis that the treatment has no effect on the observed outcome variable. According to this method, for each matched pair (t_u, t_v) a rank is assigned to the outcome difference $Y(t_u^y) - Y(t_v^y)$. The Wilcoxon's signed rank statistic W is estimated as the sum of the ranks of the positive differences (the interested reader can find a detail description of the method in [51]). Under the null hypothesis, the mean value of W is $S \cdot (S + 1)/4$, where S the number of matched samples. When S is sufficiently large, the upper bound of the distribution of W can be approximated by a normal distribution with mean $\Gamma/(1+\Gamma) \cdot S \cdot (S+1)/2$. Thus, the sensitivity on unobserved confounding variables can be assessed by computing the upper bounds on the p-values of the Wilcoxon's signed rank test for increasing Γ values.

We apply this method in order to evaluate the sensitivity of our results on unobserved confounding variables. At Table 5, we present the results of our sensitivity analysis for $\Gamma \leq 2.0$ and for the $T2$ pair of thresholds. According to our results, the causal influence of Twitter on Apple stock prices would be considered statistically significant (with p-value 0.014) even if some days were twice more likely (i.e. $\Gamma = 2$) to have positive sentiment conditional to the observed confounding variables due to unmeasured factors. Similarly, for Amazon our causal inference results are statistically significant (i.e. p-value < 0.05) for $\Gamma \leq 1.9$, for Yahoo! for $\Gamma \leq 1.7$ and finally for Microsoft our conclusion would be invalid for $\Gamma \geq 1.6$.

5. Related work

Several works have previously examined the potential of information extracted from social media, search engine query data or other web-related information to predict stock market returns. For example, in [8] the authors demonstrate that the level of optimism/pessimism, which is estimated using Twitter data, correlates with stock market movement. Other projects have been focussed on the possible use of sentiment analysis based on Twitter data for the prediction of traded assets prices by applying a bivariate Granger causality analysis [3,9] or regression models [52]. Similarly, in [53] information theoretic

Table 5
Sensitivity analysis.

Γ	Upper bound on p-value			
	AAPL	AMZN	YHOO	MSFT
1.0	0.0000	0.0000	0.0000	0.0000
1.1	0.0000	0.0000	0.0000	0.0000
1.2	0.0000	0.0000	0.0000	0.0005
1.3	0.0000	0.0002	0.0000	0.0027
1.4	0.0001	0.0011	0.0003	0.0109
1.5	0.0003	0.0042	0.0010	0.0327
1.6	0.0009	0.0131	0.0034	0.0775
1.7	0.0021	0.0328	0.0091	0.1519
1.8	0.0043	0.0692	0.0209	0.2552
1.9	0.0082	0.1263	0.0419	0.3789
2.0	0.0142	0.2050	0.0749	0.5093

methods are used to investigate whether sentiment analysis of social media can provide statistically significant information for the prediction of stock markets and in [54] authors propose a prediction method based on machine learning. In [7,55] the authors have also demonstrated that search engine query data correlate with stock market movements. In [56] the authors propose a trading strategy that utilizes information about Wikipedia views. They demonstrate that their trading strategy outperforms random strategy. However, all the above mentioned studies are based on bivariate models. Although their results indicate that social media and other web sources may carry useful information for stock market prediction, by using these techniques it is not possible to figure out if other factors are influencing the observed trends.

Trading strategies that utilize both technical analysis and sentiment analysis are discussed in [57,58]. However, these works are based on regression analysis, thus they suffer from the limitations that have been previously discussed. Moreover, all the studies so far focus mainly on prediction of stock market movement. Although they provide insights about the influence that emotional and social factors may have on stock market, they do not investigate the presence of causality. To the best of our knowledge, this is the first work that attempts to measure the causal effect of such factors on stock markets.

6. Discussion

Our results on the simulated experiments indicate that our method is more effective on avoiding false positive causality conclusions. We have examined the performance of the proposed method in datasets with up to 50 dimensions and 100 time-samples. Extreme high-dimensional cases with $p > n$ are not considered in this study. In such cases, balancing treatment and control groups for each confounding variable would require large number of samples. *Propensity score matching* [26] represents an alternative matching method that can effectively handle large number of confounding variables by performing matching on a single balancing score, i.e. the *propensity score*. The propensity score corresponds to the probability of a unit to be assigned to a treatment and it is usually approximated by applying a logistic regression model of the treatment against the set of confounding variables. *High-dimensional propensity score matching* [59] has also been proposed in order to handle extreme high-dimensional cases with $p \gg n$.

One of the main advantages of the proposed MDT method over multivariate Granger causality and CMI is that the design of the study is separated from the analysis. The values of the response time-series Y are not used during the matching process. The causal impact of a time-series X on Y is evaluated only after sufficient balance between the treated and untreated samples has been achieved. In contrast, in a regression-based analysis the response time-series Y is used in order to learn the coefficients of the predictor variables of the study. Many studies suggest that regression-based methods for causal inference are less reliable [60]. Moreover, the proposed method is non-parametric, while Granger causality is based on assumptions about the model class (i.e. linear/non-linear relationships). According to our results, linear Granger causality performs poorly when there are non-linear relationships among the examined time-series.

In addition, as it was previously discussed, MDT requires significantly fewer conditional independence tests and smaller conditioning sets. In detail, the maximum conditioning set of the proposed method is equal to the maximum lag L, while the maximum conditioning set of CMI is $M \cdot L$ (with M the number of confounding variables). Thus, MDT can handle more effectively datasets which include large number of confounding variables.

Moreover, the computational cost of creating the graph is significantly lower for MDT compared to CMI. Assuming discrete time-series with values in a set V, the computational complexity of creating a graph by applying the proposed method is $O(|V|^L \cdot M \cdot N)$, while the computational cost of CMI is $O(|V|^{M \cdot L} \cdot M \cdot N)$, with $|V|$ the size of set V. However, causal inference with MDT requires an additional matching step, and consequently, its computational complexity largely depends on the matching method that is applied. If a simple nearest neighbor matching method is applied [16], the cost of finding the best match of a single unit is $O(|\mathbf{H}| \cdot N)$, with $|\mathbf{H}|$ denoting the size of set $\mathbf{H}$, and the cost of matching all the units is $O(|\mathbf{H}| \cdot N^2)$. Thus, the overall computational cost of MDT is $O(|V|^L \cdot N + |\mathbf{H}| \cdot N^2)$. However, when more complex matching methods, such as Genetic matching [46], are applied, the computational cost of MDT can be significantly larger. Genetic matching algorithm applies an evolutionary search method in order to find optimal weights for each covariate. In each algorithm iteration, a set of P weights for each one of the variables in set $\mathbf{H}$ is generated. P corresponds to the *population size* of the genetic algorithm.

Then, nearest neighbor matching is applied on the weighted variables of **H** for each one of the P weights. A loss function is used to estimate the loss for each of the P resulted sets of matched pairs. If the loss is sufficiently small for any of the P weights, the method terminates; otherwise this process is repeated. A maximum number of iterations I can be used in order to set an upper bound on the computational time of the method. In our experiments, we used as loss function the average standardized mean difference between the treated and control units. The cost of loss estimation for each weights set is $O(|\mathbf{H}| \cdot N)$. Thus, the total computational cost of the matching process is $O(I \cdot P \cdot |\mathbf{H}| \cdot N^2)$. Although the computational cost of MDT could be significantly larger than the cost of CMI, given the availability of advanced computational resources, the computational efficiency can be traded for more reliable results. In our simulated experiments, the running time of CMI was in order of seconds while the running time of MDT was in order of minutes, using a 2.6 GHz quad core CPU and 16 GB RAM.

Finally, we discuss the assumptions behind our method, thus outlining situations where the method is expected to perform well. There are four key ingredients in our method:

1. We perform independence tests on time series pairs. There is no guarantee that if there were higher-order dependencies among several time-series (e.g. the outcome variable and two confounding variables), they would be detected by the pair-wise tests.
2. In our conditional independence testing, the maximum conditioning set is determined by the largest lag L. The value of L is of course upper bounded by our desire to have sample sizes large enough to yield sufficient power to independence tests.
3. The matching procedure assumes that there is an overlap in the confounding variables' values between the groups of treated and control units. If this is not the case, the matching will not achieve sufficient balance. For example, in our case, it was not feasible to conduct a causality study of the impact of social media sentiment on the treated assets prices of Apple Inc. When the $T3$ thresholds pair is used, since there was not sufficient overlap on the confounding variables' values.
4. The estimation of the average treatment effect is influenced by the power of the statistical test that is applied.

7. Conclusions

In this study, for the first time we have attempted to quantify the causal impact of social and emotional factors, captured by social media, on daily stock market returns of individual companies (i.e., not just a mere correlation between the two). We have proposed a novel non-parametric framework for causal analysis in time-series. Our evaluation on synthetic data demonstrates that our method is more effective on inferring true causality and avoiding false positive conclusions compared to other methods that have been previously used for causal inference in time-series. Our approach can incorporate a large number of factors and, therefore, can effectively handle complex data such as financial data. Indeed, causality studies that are based on observational data rather than experimental procedures could be biased in case of missing confounding variables. However, conducting experimental studies is not feasible in most cases. In this work we have minimized the risk of biased conclusions due to unmeasured confounding variables by including in our study a large number of factors. Additionally, we conduct an analysis on the sensitivity of our conclusions on missing confounding variables. We have estimated a sentiment index indicating the probability that the general sentiment of a day, based on tweets posted for a target company, is positive. Our results show that Twitter data polarity does indeed have a causal impact on the stock market prices of the examined companies. Although our study involves only big technological companies and consequently our findings might not be generalizable, especially for companies that do not focus directly on retail customers. The analysis of the validity of these findings in the B2B sector is an open question that we plan to explore. Hence, we believe social media data could represent a valuable source of information for understanding the dynamics of stock market movements.

References

[1] Sitaram Asur, Bernardo A. Huberman, Predicting the future with social media, in: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, (WI-IAT'10), IEEE, 2010, pp. 492–499.
[2] Andranik Tumasjan, Timm Oliver Sprenger, Philipp G. Sandner, Isabell M. Welpe, Predicting elections with Twitter: What 140 characters reveal about political sentiment, in: Proceedings of the 4th International Conference on Weblogs and Social Media (ICWSM'10), volume 10, 2010, pp. 178–185.
[3] Johan Bollen, Huina Mao, Xiaojun Zeng, Twitter mood predicts the stock market, J. Comput. Sci. 2 (1) (2011) 1–8.
[4] R.M. Bond, et al., A 61-million-person experiment in social influence and political mobilization, Nature 489 (7415) (2012) 295–298.
[5] L. Muchnik, S. Aral, S.J. Taylor, Social influence bias: A randomized experiment, Science 341 (6146) (2013) 647–651.
[6] John Concato, Nirav Shah, Ralph I. Horwitz, Randomized, controlled trials, observational studies, and the hierarchy of research designs, N. Engl. J. Med. 342 (25) (2000) 1887–1892.
[7] Tobias Preis, Daniel Reith, H. Eugene Stanley, Complex dynamics of our economic life on different scales: insights from search engine query data, Phil. Trans. R. Soc. A 368 (1933) (2010) 5707–5719.
[8] Xue Zhang, Hauke Fuehres, Peter A. Gloor, Predicting stock market indicators through Twitter "I hope it is not as bad as I fear", in: Proceedings of the 2nd Collaborative Innovation Networks Conference, Vol. 26, Elsevier, 2011, pp. 55–62.
[9] Xue Zhang, Hauke Fuehres, Peter A. Gloor, Predicting asset value through Twitter buzz, in: Advances in Collective Intelligence, Springer, 2012, pp. 23–34.
[10] Clive W.J. Granger, Some recent development in a concept of causality, J. Econometrics 39 (1) (1988) 199–211.
[11] Huina Mao, Scott Counts, Johan Bollen, Predicting financial markets: Comparing survey, news, twitter and search engine data. arXiv preprint arXiv:1112.1051, 2011.

[12] Dror Y. Kenett, Michele Tumminello, Asaf Madi, Gitit Gur-Gershgoren, Rosario N. Mantegna, Eshel Ben-Jacob, Dominating clasp of the financial sector revealed by partial correlation analysis of the stock market, PLoS One 5 (12) (2010) e15032.
[13] K. Tse Chi, Jing Liu, Francis C.M. Lau, A network perspective of the stock market, J. Empir. Finance 17 (4) (2010) 659–667.
[14] Jianqing Fan, Fang Han, Han Liu, Challenges of big data analysis, Natl. Sci. Rev. 1 (2) (2014) 293–314.
[15] William R. Shadish, Thomas D. Cook, Donald Thomas Campbell, Experimental and Quasi-Experimental Designs for Generalized Causal Inference, Wadsworth Cengage learning, 2002.
[16] E.A. Stuart, Matching methods for causal inference: A review and a look forward, Stat. Sci.: Rev. J. Inst. Math. Stat. 25 (1) (2010) 1.
[17] P.W. Holland, Statistics and causal inference, J. Amer. Statist. Assoc. 81 (396) (1986) 945–960.
[18] Stephen L. Morgan, Christopher Winship, Counterfactuals and Causal Inference, Cambridge University Press, 2014.
[19] Peter M. Bentler, Multivariate analysis with latent variables: Causal modeling, Annu. Rev. Psychol. 31 (1) (1980) 419–456.
[20] Paul R. Rosenbaum, Donald B. Rubin, Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome, J. R. Stat. Soc. Ser. B Stat. Methodol. (1983) 212–218.
[21] David Freedman, From association to causation via regression, Adv. Appl. Math. 18 (1) (1997) 59–110.
[22] Guido W. Imbens, Nonparametric estimation of average treatment effects under exogeneity: A review, Rev. Econ. Stat. 86 (1) (2004) 4–29.
[23] Peter C. Austin, The relative ability of different propensity score methods to balance measured covariates between treated and untreated subjects in observational studies, Med. Decis. Making 29 (6) (2009) 661–677.
[24] Peter C. Austin, Goodness-of-fit diagnostics for the propensity score model when estimating treatment effects using covariate adjustment with the propensity score, Pharmacoepidemiol. Drug Safety 17 (12) (2008) 1202–1217.
[25] Peter C. Austin, Assessing balance in measured baseline covariates when using many-to-one matching on the propensity-score, Pharmacoepidemiol. Drug Safety 17 (12) (2008) 1218–1225.
[26] Valerie S. Harder, Elizabeth A. Stuart, James C. Anthony, Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research, Psychol. Methods 15 (3) (2010) 234.
[27] Judea Pearl, An introduction to causal inference, Int. J. Biostat. 6 (2) (2010).
[28] Judea Pearl, Causality: Models, Reasoning and Inference, Cambridge University Press, 2009.
[29] Judea Pearl, Causal diagrams for empirical research, Biometrika 82 (4) (1995) 669–688.
[30] Markus Kalisch, Peter Bühlmann, Estimating high-dimensional directed acyclic graphs with the pc-algorithm, J. Mach. Learn. Res. 8 (Mar) (2007) 613–636.
[31] Junning Li, Z. Jane Wang, Controlling the false discovery rate of the association/causality structure learned with the pc algorithm, J. Mach. Learn. Res. 10 (Feb) (2009) 475–514.
[32] Peter Spirtes, An anytime algorithm for causal inference, in: AISTATS, Citeseer, 2001.
[33] Diego Colombo, Marloes H. Maathuis, Markus Kalisch, Thomas S. Richardson, Learning high-dimensional directed acyclic graphs with latent and selection variables, Ann. Statist. (2012) 294–321.
[34] Clive W.J. Granger, Investigating causal relations by econometric models and cross-spectral methods, Econometrica (1969) 424–438.
[35] Adam B. Barrett, Lionel Barnett, Anil K. Seth, Multivariate granger causality and generalized variance, Phys. Rev. E 81 (4) (2010) 041907.
[36] Kun Zhang, Jonas Peters, Dominik Janzing, Bernhard Schölkopf, Kernel-based conditional independence test and application in causal discovery. arXiv preprint arXiv:1202.3775, 2012.
[37] Jakob Nawrath, M. Carmen Romano, Marco Thiel, István Z. Kiss, Mahesh Wickramasinghe, Jens Timmer, Jürgen Kurths, Björn Schelter, Distinguishing direct from indirect interactions in oscillatory networks with multiple time scales, Phys. Rev. Lett. 104 (3) (2010) 038701.
[38] Jonas Peters, Dominik Janzing, Bernhard Schölkopf, Causal inference on time series using restricted structural equation models, in: Advances in Neural Information Processing Systems, 2013, pp. 154–162.
[39] Doris Entner, Patrik O. Hoyer, On causal discovery from time series data using fci, Probab. Graph. Models (2010) 121–128.
[40] Thomas Schreiber, Measuring information transfer, Phys. Rev. Lett. 85 (2) (2000) 461.
[41] Lionel Barnett, Adam B. Barrett, Anil K. Seth, Granger causality and transfer entropy are equivalent for gaussian variables, Phys. Rev. Lett. 103 (23) (2009) 238701.
[42] Bernd Pompe, Jakob Runge, Momentary information transfer as a coupling measure of time series, Phys. Rev. E 83 (5) (2011) 051122.
[43] Jakob Runge, Jobst Heitzig, Vladimir Petoukhov, Jürgen Kurths, Escaping the curse of dimensionality in estimating multivariate transfer entropy, Phys. Rev. Lett. 108 (25) (2012) 258701.
[44] Bo Lu, Elaine Zanutto, Robert Hornik, Paul R. Rosenbaum, Matching with doses in an observational study of a media campaign against drug abuse, J. Amer. Statist. Assoc. 96 (456) (2001) 1245–1253.
[45] Keisuke Hirano, Guido W. Imbens, The propensity score with continuous treatments, in: Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives, 2004, pp. 73–84. 226164.
[46] Alexis Diamond, Jasjeet S. Sekhon, Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies, Rev. Econ. Stat. 95 (3) (2013) 932–945.
[47] Jasjeet S. Sekhon, Opiates for the matches: Matching methods for causal inference, Ann. Rev. Political Sci. 12 (2009) 487–508.
[48] Donald B. Rubin, Estimating causal effects from large data sets using propensity scores, Ann. Intern. Med. 127 (8_Part_2) (1997) 757–763.
[49] Mike Thelwall, Heart and soul: Sentiment strength detection in the social web with sentistrength, Cyberemotions (2013) 1–14.
[50] Paul R. Rosenbaum, Observational studies, in: Observational Studies, Springer, 2002, pp. 1–17.
[51] Erich Leo Lehmann, H.J. D'abrera, Nonparametrics: Statistical Methods Based on Ranks, Holden-Day, 1975.
[52] Jianfeng Si, Arjun Mukherjee, Bing Liu, Qing Li, Huayi Li, Xiaotie Deng, Exploiting Topic Based Twitter Sentiment for Stock Prediction, in: ACL (2), Vol. 2, 2013, pp. 24–29.
[53] Ilya Zheludev, Robert Smith, Tomaso Aste, When can social media lead financial markets? Sci. Rep. 4 (2014).
[54] Alexander Porshnev, Ilya Redkin, Alexey Shevchenko, Machine learning in prediction of stock market indicators based on historical data and data from twitter sentiment analysis, in: 2013 IEEE 13th International Conference on Data Mining Workshops, IEEE, 2013, pp. 440–444.
[55] Tobias Preis, Helen Susannah Moat, H. Eugene Stanley, Quantifying trading behavior in financial markets using Google Trends, Sci. Rep. 3 (2013).
[56] Helen Susannah Moat, Chester Curme, Adam Avakian, Dror Y. Kenett, H.Eugene Stanley, Tobias Preis, Quantifying wikipedia usage patterns before stock market moves, Sci. Rep. 3 (2013).
[57] Robert P. Schumaker, Hsinchun Chen, Textual analysis of stock market prediction using breaking financial news: The azfin text system, ACM Trans. Inf. Syst. (TOIS) 27 (2) (2009) 12.
[58] Shangkun Deng, Takashi Mitsubuchi, Kei Shioda, Tatsuro Shimada, Akito Sakurai, Combining technical analysis with sentiment analysis for stock price prediction, in: Proceedings of IEEE Ninth International Conference on Dependable, Autonomic and Secure Computing, (DASC'11), IEEE, 2011, pp. 800–807.
[59] Sebastian Schneeweiss, Jeremy A. Rassen, Robert J. Glynn, Jerry Avorn, Helen Mogun, M. Alan Brookhart, High-dimensional propensity score adjustment in studies of treatment effects using health care claims data, Epidemiol. (Cambridge, Mass.) 20 (4) (2009) 512.
[60] Peter C. Austin, An introduction to propensity score methods for reducing the effects of confounding in observational studies, Multivariate Behav. Res. 46 (3) (2011) 399–424.