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

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,