Signal Processing 124 (2016) 266–278

Contents lists available at ScienceDirect

Signal Processing

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

On characterizing scale effect of Chinese mutual funds via text mining

Heng Wang [a], Junjie Wu [b,*], Shi Yuan [b], Jian Chen [a]

[a] School of Economics and Management, Tsinghua University, Beijing 100084, China
[b] School of Economics and Management, Beihang University, Beijing 100191, China

ARTICLE INFO

Article history:
Received 21 February 2015
Received in revised form
23 May 2015
Accepted 28 May 2015
Available online 14 June 2015

Keywords:
Text mining
K-means
Mutual fund
Scale effect

ABSTRACT

This paper investigates the correlation between mutual funds' scale and return in China by text mining on the sheer volume of online financial reports. We crawl the webpages of all Chinese open-end mutual funds from a well-known financial website, which are then parsed to obtain time-series data of fund scales and returns. We argue that with long-tail distribution of fund scales, to examine the correlation directly in an individual level is not appropriate; rather, we should consider it in a group level by scales and take different market conditions into consideration. To illustrate this, we start with a data-fitting test to demonstrate that the tail of fund scale fits best in a distribution between Power-Law and Log-Normal. Hence, to categorize mutual funds by equal scale could lead to fund groups in substantially different sizes, and the subsequent results are thus prone to bias. We therefore introduce *K*-means clustering for fund categorization, which enables reliable examination of correlations between fund scale and return. Empirical study unveils some interesting findings on the scale effect of funds under different market conditions. These findings highlight the uniqueness of emerging markets while providing interesting guidelines for exploiting big data analytics for financial studies.

1. Introduction

Recent decades have witnessed the booming of financial markets in emerging countries. However, little is known so far about whether these markets have followed the footprints of financial markets of developed countries. This paper aims to examine whether classic theories for financial market still hold for emerging countries. Specifically, we choose to test the correlation between fund scale and the performance with the important help of text retrieval on the sheer volume of online financial reports. This is not unusual since big data analytics has been widely adopted in various application domains [8,17,21].

Previous studies have reported mixed findings, most of which are with data from financial markets of developed countries. For example, it is believed that, fund scale increases at the cost of reducing liquidity and increasing administrative expenses, which may result in worse fund performance [2,13,16]; on the contrary, positive impacts, such as economic of scale, are also identified in the mutual fund market [18]. Interestingly, there is also a line of literature claiming no significant correlations [1,12,5,9]. Besides, some researchers have proposed a threshold-based structure: The cost corresponding to a larger fund scale may increase dramatically only when it is beyond a threshold [22,14,4], such that the performance is associated with the scale via an inverse-U relationship.

The financial markets in emerging economies, such as China's market examined in this paper, usually differ from those in developed countries due to their short lifecycle and unprecedented growth. This indeed distinguishes our research

* Corresponding author.
E-mail addresses: 1923878540@qq.com (H. Wang), wujj@buaa.edu.cn (J. Wu), ystone1025@163.com (S. Yuan), jchen@mail.tsinghua.edu.cn (J. Chen).

http://dx.doi.org/10.1016/j.sigpro.2015.05.018
0165-1684/

from the literature on financial markets in developed countries. In China, the fund market started as late as in 1997, while the first mutual fund was not established until 2001. As a result, for the risk concern, the regulatory obligation on mutual funds was ever too tight compared with the regulation in developed countries. For instance, the investment instruments did not include commercial papers, real estate and precious metals; the market entry barriers were also too high for fund managers and fund trustees, which prevented the market from free competition. In June 2013, the Chinese government issued the new Law on Funds for Investment in Securities, which reflects the spirit of deregulation of mutual funds, but is yet subject to the operability problem. Nevertheless, the mutual fund market in China quickly reached a scale of 6.94 trillion Yuan by the end of 2006, representing 81% of the China's fund market.

Since 2006, China's mutual fund market has gone through various market conditions including the financial crisis and a series of financial stimulus and recovery plans. As a result, data from China's mutual fund market are extremely fruitful for researcher to examine the scale effect of mutual funds. However, due to the data availability on limited time horizon, previous studies on the China's mutual fund market have not reached an agreement on the correlations between fund scale and return. Positive scale effects are discovered in the closed-end fund market [30]. Other studies also identify no strong correlations between the scale and return [29]. Since the long tail of fund scale would lead to model complexity in examining the scale effects of fund returns, the classic beta-portfolio method introduced by Fama and French [6,7] is widely adopted among existing literature.

Fig. 1 shows intuitively the relationship between the average scales and abnormal returns of 815 mutual funds selected from China's market across 5 years (2006–2011). As can be seen, from individual fund level, the scale effect is two-folded. On one hand, large-scale funds do not show positive scale effect with approximately zero weekly abnormal return. On the other, small funds diverse in their returns, with a dozen having relatively high abnormal returns but most suffering from negative returns. These observations, however, are suspectable without considering two important factors. The first factor is the market condition, under which the scale effect may have different manifestations with varied significance levels. For instance, China's stock market experienced clear market-rise, decline and fluctuation stages from 2006 to 2011, and hence we cannot discuss the scale effect of funds simply over whole time period. The second factor is fund grouping, which suggests observing scale effect of funds from a group rather than individual level. This strategy, although very simple in essence, may help to shield the interferences and make us focus on the mainstream rules.

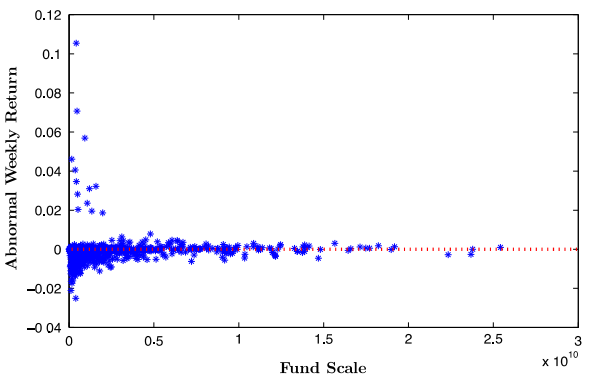

Fig. 1. Relationship between fund scale and return in an individual level.

Motivated by the above considerations, this paper examines the scale effect of China's fund market with a more up-to-date dataset covering various market conditions, which enables us to test the comprehensiveness and robustness of our findings. This dataset is obtained by searching, crawling and parsing the sheer volume of webpages about all the Chinese open-end mutual funds published on a well-known financial website: EastMoney (http://www.eastmoney.com/). To our best knowledge, this paper is among the earliest research on leveraging online textual big data for financial studies of Chinese market. Based on this dataset, we explore the long-tail distribution of fund scale and its impact to the fund return. Specifically, to overcome the bias induced by the long tail of fund scale, we employ a fund grouping strategy to unveil the scale and return relationship that was otherwise ambiguous in an individual level.

To start with, we character the specific distribution of the long tail of fund scale, which turns out to be the Power-Law distribution in the steady state in theory. We then examine the tail part of fund scale empirically by comparing two distributions, i.e., the Power-Law distribution [19,23,10] and the Log-Normal distribution [24], both of which are widely adopted in literature for data fitting of mutual fund scale distribution. Interestingly, we discover that a hybrid distribution between the Power-Law and Log-Normal distributions actually fits best in the tail part of scale distribution of China's mutual funds. This suggests that the long-tail effect of mutual funds in emerging market such as in China is not as significant as in developed markets.

To deal with the long tail effect of fund scale, we further introduce *K*-means clustering algorithm [20,11] for fund categorization, and examine the correlations between fund scale and return in a group level. Indeed, recent studies have demonstrated the benefits from employing pattern recognition methods for financial engineering [15,26]. We choose *K*-means since it can group funds naturally by scale and thus avoid the latent bias induced by categorizing funds evenly or in percentiles with a long tail. Our results suggest that, (1) the conventional belief that smaller funds receive higher return is only valid during market decline; (2) during market rise, the fund scale is positively correlated with weekly abnormal return; (3) during market fluctuation, however, an inverse-U relationship appears with weekly abnormal return, which indicates the existence of the optimum scale. These findings highlight the uniqueness of emerging markets while providing interesting guidelines for future research.

The rest of the paper is organized as follows. Section 2 introduces the theoretical framework, based on which we fit our data into both Power-Law and Log-Normal distributions in Section 3. Section 4 empirically studies the correlation between fund scale and return. Section 5 discusses our findings and concludes.