Electronic Commerce Research and Applications 14 (2015) 444–455

Contents lists available at ScienceDirect

Electronic Commerce Research and Applications

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

Concept drift mining of portfolio selection factors in stock market

Yong Hu [a,1], Kang Liu [b,1], Xiangzhou Zhang [c,a,1], Kang Xie [c,*], Weiqi Chen [d], Yuran Zeng [b], Mei Liu [e,*]

[a] Big Data Decision Institute, Jinan University, Guangzhou, PR China
[b] School of Management, Guangdong University of Foreign Studies, Guangzhou, PR China
[c] School of Business, Sun Yat-sen University, Guangzhou, PR China
[d] Faculty of Automation, Guangdong University of Technology, Guangzhou, PR China
[e] Department of Internal Medicine, Division of Medical Informatics, University of Kansas Medical Center, Kansas City, KS 66160, USA

ARTICLE INFO

Article history:
Received 25 November 2014
Received in revised form 19 May 2015
Accepted 14 June 2015
Available online 8 July 2015

Keywords:
Concept drift mining
Stock analysis
Cross-sectional analysis
Causal discovery
Modified Additive Noise Model with Conditional Probability Table
China stock market

ABSTRACT

Concept drift is a common phenomenon in stock market that can cause the devaluation of the knowledge learned from cross-sectional analysis as the market changes over time in unforeseen ways. The widely used cross-sectional regression analysis based on expert knowledge has obvious limitations in handling problems that involve concept drift and high-dimensional data. To discover causal relations between portfolio selection factors and stock returns, and identify concept drifts of these relations, we apply a novel causal discovery technique called modified Additive Noise Model with Conditional Probability Table (ANMCPT). In evaluation experiments, we compares ANMCPT to the conventional cross-sectional analysis approach (i.e., Fama–French framework) in mining relationships between portfolio selection factors and stock returns. Results indicate that the factors selected by ANMCPT outperform the factors adopted in most previous cross-sectional researches that followed the Fama–French framework. To the best of our knowledge, this paper is the first to compare causal inference technique with Fama–French framework in concept drift mining of stock portfolio selection factors. Our causal inference-based concept drift mining method provides a new approach to accurate knowledge discovery in stock market.

1. Introduction

Stock market has always been followed closely by investors all over the world; electronic order book trade in principal markets has reached 48 trillion USD in 2012 (World Federation of Exchanges 2013). However, stock trading is accompanied with great risk, and all market participants strive to trade with higher risk-adjusted returns (Atsalakis and Valavanis 2009, Bahra mmirzaee 2010, Bodie et al. 2005, Fama and French 1992, Malkiel and Fama 1970). One of the key challenges to successful stock market investment is the accurate and timely recognition of informative factors that are closely tied to the expected returns and risks of stocks (called effective portfolio selection factors), as well as the (causal) relationships between these factors and the future returns and risks of stocks.

Concept drift/shift is a phenomenon that the underlying data distribution changes over time, space and conditions, making the relationships between variables found in the past inconsistent with the new data (Delany et al. 2005, Tsymbal 2004, Wang et al. 2003, Zliobaite 2009). According to the efficient market theory (Malkiel and Fama 1970) and numerous empirical researches (Chan et al. 2000, Fama and French 1998, Lam 2002, Lucas et al. 2002, Wang and Di Iorio 2007), effective portfolio selection factors in stock market would change over time and markets; therefore, timely identification of this concept drift is a key problem to stock investment.

Existing researches provide an incomplete view of the concept drift issue in stock market due to limitations in analysis approach and factor selection. Cross-sectional analysis is a conventional approach to discover effective portfolio selection factors, and the most famous one is the cross-sectional regression framework presented in (Fama and French 1992) (hereafter Fama–French framework). The Fama–French framework selects factors according to previous researches and examines the effect of those factors on stock returns using long-term data and multivariable cross-sectional regression model. Although the Fama–French framework were used in many studies and have identified numerous effective factors in various markets, several limitations of this

* Corresponding authors at: School of Business, Sun Yat-sen University, Guangzhou, PR China (K. Xie), and Department of Internal Medicine, Division of Medical Informatics, University of Kansas Medical Center, Kansas City, KS 66160, USA (M. Liu).

E-mail addresses: henryhu200211@163.com (Y. Hu), researcher_lk@foxmail.com (K. Liu), zhxzhou86@foxmail.com (X. Zhang), mnsxk@mail.sysu.edu.cn (K. Xie), isscwqcxz@126.com (W. Chen), zengyuran@hotmail.com (Y. Zeng), meiliu@kumc.edu (M. Liu).

[1] Co-first authors.

http://dx.doi.org/10.1016/j.elerap.2015.06.002
1567-4223/

approach were observed: First, it is ineffective in handling high dimensional data, because testing all possible regression models of candidate factors is often complicated and time consuming, and thus sufficient priori knowledge is needed for feature selection. Second, the selection of factors is theoretically not optimal because most fundamental researches were about US market before 1990; applying factors considered in these researches to other markets might suffer from a concept drift problem. Lastly, it might neglect the nonlinear effects between factors, and, most importantly, regression model identifies correlations rather than causalities.

Compared to the conventional cross-sectional analysis approach, combining causal inference and concept drift adaptation techniques is a more promising approach to discover causalities from high dimensional stock market data. Prior researches have proposed causal discovery techniques for two-to-one and one-to-one causality (Hoyer et al. 2008, Hu et al. 2013, Liu et al. 2014, Zhang et al. 2014). However, the number of factors that influence stocks is uncertain. Thus, this paper proposes a novel causal discovery technique called modified Additive Noise Model with Conditional Probability Table (ANMCPT) to address the problem of many-to-one causality discovery.

In this study, we collected cross-sectional data from China stock market, covering a 13-year period of July 1999 to June 2011. To evaluate the validness and effectiveness of ANMCPT, we first applied the conventional Fama–French framework on a low-dimensional data (Dataset I). Both vertical and horizontal drifts of the relations between factors and stock returns and of the relations among factors were observed. Then, we applied ANMCPT to the same dataset (Dataset I). Results showed that ANMCPT can produce consistent result. Finally, we applied ANMCPT to a high-dimensional data (Dataset II, which consists of 53 factors) and conducted a concept drift analysis. Results revealed obvious concept drifts–the most informative factors changed over time. Moreover, the factors selected by ANMCPT outperformed those by Fama–French framework when being used to construct stock prediction models. This demonstrates the importance of applying causal discovery technique for concept drift mining when we conducts cross-sectional analysis.

The contributions of this paper could be concluded as two points. First, in contrast to most existing approaches for cross-sectional analysis such as the Fama–French framework, our method can not only handle many-to-one causality discovery in high-dimensional dynamic stock markets, but also empirically outperform the classic Fama–French framework. Second, we have clearly exhibited and analyzed the concept drift phenomenon of effective portfolio selection factors. To the best of our knowledge, this paper is the first to compare causal inference technique with Fama–French framework in mining concept drifts of effective portfolio selection factors. The proposed method will provide a new approach and framework for accurate knowledge discovery in stock markets.

The remainder of this paper is organized as follows. Section 2 reviews the related works on cross-sectional analysis and concept drift. Section 3 introduces the Fama–French framework and the ANMCPT method. Section 4 describes two datasets used in our experiments while Section 5 presents the comparative empirical results and analyses. The last section provides conclusions.

2. Literature review

2.1. Cross-sectional analysis

Cross-sectional analysis is a classical investment analysis method. It is different from pattern recognition and time series analysis which are used in technical analysis to identify patterns of price movements of a single stock. Cross-sectional analysis devotes to search for factors that can explain the differences of return between various stocks (Fama and French 1992, 1998; Wang and Xu 2004 and Wang and Di Iorio 2007). Researchers have confirmed many well-known effective portfolio selection factors based on the long-term data (always more than ten years) of a wide range of stocks (such as all stocks in the market, mainly in US stock markets), such as earnings-price ratio (E/P), firm size and book-to-market equity (B/P).

Most effective portfolio selection factors examined by former researches are calculated according to the financial statements of a listed company. However, it doesn't mean that investors who buy a stock having a bad operating situation will definitely lose or vice versa. According to the efficient market theory, the effectiveness of markets will influence the effects of different factors on future stock returns (Malkiel and Fama 1970). On one extreme, if a market is a perfect market, all existing information will be rationally and instantly reflected in stock price, then no investors can gain excess return over the market average, and the differences in return between diverse portfolios are only related to their differences in risk (usually measured by Beta). For example, the Capital Asset Pricing Model (CAPM), established by Sharpe (1964), Lintner (1965) and Black (1972), is such a model that describes the relationship between Beta and expected return of stocks. And it suggests that using Beta is sufficient to describe the cross-sectional differences in expected returns. This idea was supported by (Fama and MacBeth 1973).

However, several subsequent empirical studies challenged CAPM and turned to explore other effective factors for explaining the cross-section of expected stock return. For example, Banz (1981) discovered that smaller firms have a higher average risk–adjusted return than the large firms (Size effect). Although CAPM implies that the effect of leverage can be captured by Beta, Bhandari (1988) found that by controlling the beta and firm size, positive relation exists between leverage and average return. Stattman (1980) and Rosenberg et al. (1985) confirmed the positive relation between expected stock return and book-to-market equity (B/P). Basu (1983) claimed that stocks with higher earnings-price ratio (E/P) earn higher average risk-adjusted return when firm size is controlled.

It is reasonable to anticipate that the effects of the above factors may be partly redundant or decrease when considering other factors. For example, Ball (1978) suggested that E/P is a catch-all proxy for unnamed factors in expected returns. Thus, Fama and French (1992) examined the joint effect of these factors in US stock market of the period between 1962 and 1990, and found that Beta does not help explain the cross-section of average stock returns and the combination of size and B/P can (seems able to) capture the effect of leverage and E/P on average stock returns. Later research by Fama and French (1996) has also considered the effect of cash flow/price (C/P). Many researches have followed Fama and French's approach to investigate these factors in other principal markets (Chan et al. 2000, Fama and French 1998, Lam 2002, Lucas et al. 2002, Wang and Di Iorio 2007).

2.2. Concept drift in stock market analysis

Concept drift indicates the situation where the underlying data distribution change over time, space and condition; thus the relationships between variables found in the old data become inconsistent or irrelevant in the new data (Delany et al. 2005, Tsymbal 2004, Wang et al. 2003, Zliobaite 2009). Concept drift problem can be observed in stock markets when comparing results of existing cross-sectional analysis researches (Chan et al. 2000, Fama and French 1998, Lam 2002, Lucas et al. 2002, Wang and Di Iorio 2007)

as listed in Table 1. Previous researches showed that concept drifts in many factors (such as economic growth, level of market liberalization, and maturity of the institutions, regulations, laws and the investors) will cause concept drifts in the relationships between stock returns and factors (Bilson et al. 2001, Chan et al. 2000, Filis 2010, Gay 2008, Henry 2000, Lucas et al. 2002, Morck et al. 2000, Wang and Xu 2004, Wang and Di Iorio 2007). In general, concept drifts can be categorized as vertical (time-dimensional) drift and horizontal (space-dimensional or market-dimensional) drift.

For vertical drift, validity of relationship may decrease over time, and causes unexpected risks to investors. For example, researchers have verified that size effect and value premium (i.e., the size, B/P and E/P effects on returns) existed in almost all the global mature markets during the period of 1974–1994 (Fama and French 1992, 1998). However, the effects of these factors on stock returns have significantly changed over time (Chan et al. 2000, Lucas et al. 2002, Malkiel and Fama 1970). Chan et al. (2000) found that the typical size and value effects were reversed during the period from 1990s through 1980s in the US market, which suggested that future growth in profitability may differ radically from the past. Lucas et al. (2002) indicated that the effects of value and size factors were not stable over time and for some periods they departed from the long-term patterns in the US market.

For horizontal drift, the differences between stock markets may decrease the validness of the relationships between cross-sectional factors and stock returns from one market to another. Researches have compared these relationships among many kinds of markets, and empirical results have confirmed the differences (Bilson et al. 2001, Gay 2008, Henry 2000, Lam and Spyrou 2003, Lam 2002). For example, Hong Kong market differed from US market in the effects of particular accounting factors, such as firm size (Lam and Spyrou 2003, Lam 2002). Wang and Xu (2004) found that the book-to-market variable has no significant correlation with the monthly average returns during the period from July 1996 to June 2002 in China market, which indicated the existence of horizontal drift between China and US markets.

As listed in Table 1, some researches have observed both vertical and horizontal drifts over different periods and between different markets.

2.3. Computer-aided algorithm trading and concept drift mining

For more effective and efficient decision-making in the big data era, investors started to use computer algorithms for stock trading, which is called algorithmic trading (Hendershott et al. 2011). It was reported that algorithmic trading had contributed to 50% of equity trading volume in US in 2012 (The New York Times 2012). However, most existing researches are based on technical analysis, and only a very small amount of researches have considered the concept drifts in the cross-section of expected stock returns.

In technical analysis researches, data partitioning is a popular way to deal with concept drift. Specific methods of data partitioning include sliding windows (Allen and Karjalainen 1999, Garcia et al. 2010, Neely 2003), partitioning based on data similarity such as clustering analysis (Hadavandi et al. 2010, Hsu 2011, Liu et al. 2012), and partitioning based on prior knowledge such as market situations (bullish/bearish) (Gorgulho et al. 2011, Kaucic 2010, Matilla-García 2006). After partitioning the data, popular data mining techniques such as neural network, genetic algorithm and neuro-fuzzy system are used to mine the relationships between factors and stocks returns. As concept drift gains more and more attention among researchers, new techniques are developed continually. For example, Shao et al. (2014) proposed a prototype-based classification model called SyncStream to dynamically keep track of short- and long-term representative examples to capture the trends of time-changing concepts. Firstly, this model divided the selected examples into different prototypes (each prototype belongs to a class) and used them to predict new examples. Secondly, according to the predicted results, the representativeness of each prototypes would change and only the prototypes with high representativeness were reserved and clustered into a smaller set of prototypes. Thirdly, principal component analysis (PCA) and statistical analysis were used to track concept drift by evaluating the change of class distribution in various dimensions. Fourthly, previous prototypes would be adjusted when new concept was emerging. Nevertheless, Shao's method has not considered the problem of causal discovery.

In addition, although most existing data-driven studies applied data mining techniques for technical analysis, it does not mean that data mining techniques cannot play a positive role in cross-sectional analysis. In this study, we demonstrate that data mining techniques can be effective in dealing with concept drift problem in stock markets.

2.4. Causality discovery

Causal relationships are fundamental to science, and controlled randomized experiment is the basic approach to identify relations between causes and consequences (Hoyer et al. 2008, Hu et al. 2013). For many practical problems, controlled randomized experiments are either unethical, too expensive, or technically impossible (Kleinberg and Hripcsak 2011). As observational data can be collected more easily, inexpensively, or possibly than experimental data, causal relationship mining based on available observations seems a promising alternative. However, standard data analysis methods can only identify correlations rather than causalities. How to infer whether X causes Y or Y causes X from observational data is a challenging problem (Janzing et al. 2012).

For observational data-based causal inference, J. Pearl has made important contributions to both fundamental theories and practical applications (Pearl 2000, Pearl and Verma 1995). Nowadays, causal discovery techniques have been applied to many fields because of its importance and effectiveness. For example, Hu et al. (2013) proposed an algorithm called Bayesian network with causality constraints (BNCC), which is based on Bayesian networks, V-structure discovery algorithm and expert knowledge, for software project risk analysis. Mani and Cooper (2000) applied the local causal discovery algorithm to identify the causal relationships between clinical conditions and outcomes based on intensive care unit discharge summaries. In another paper, they tested a Bayesian local causal discovery algorithm in Linked Birth/Infant Death data set, and three of six uncovered relationships seemed plausible (Mani and Cooper 2004). In addition, other techniques such as additive noise models (ANM) (Hoyer et al. 2008), information geometric causal inference (Janzing et al. 2012), and causal discovery with continuous additive noise models (Peters et al. 2014) have been proposed by researchers in recent years.

3. Methodology

3.1. Fama–French framework

In order to examine relationships between cross-sectional variables and stock returns, most researchers have followed the analysis framework presented in (Fama and French 1992, Fama and MacBeth 1973). A brief introduction of this analysis method is presented as follows.

In general, the analysis process in (Fama and French 1992) is: (1) selecting factors that may affect stock returns as candidate factors; (2) conducting exploratory data analysis based on portfolios formed by sorting and grouping stocks on candidate factors; (3)

Table 1
Summary of related researches on cross-sectional analysis.

Authors	Journal	Data sources	Period	Conclusion
Fama and French (1992)	The Journal of Finance	All nonfinancial firms in the NYSE, AMEX, and NASDAQ	1962–1989	Relation between Beta and average return is also weak in the 50-year 1941–1990 period. Relations between average return and size, leverage, E/P, and book-to-market equity are strong. The negative relation between size and average return is robust to the inclusion of other variables. The positive relation between book-to-market equity and average return also persists in competition with other variables. Comparing with size effect, book-to-market equity has a consistently stronger role in average returns
Fama and French (1998)	The Journal of Finance	United States, Japan, Great Britain, France, Germany, Italy, Netherlands, Belgium, Switzerland, Sweden, Australia, Hong Kong, Singapore	1975–1995	Value stocks tend to have higher returns than growth stocks in twelve of thirteen major markets during the 1975–1995 period
Chan et al. (2000)	Financial Analysts Journal	Stocks on the NYSE, Amex, and NASDAQ	1970–1998	In 1980s, small-cap value stocks outperformance large-cap value stocks, and value stocks in general earned higher returns than growth stocks. In 1990s, large-cap stocks outperformance small-cap stocks, and value stocks were outpaced by growth stocks
Lucas et al. (2002))	Journal of Empirical Finance	Stocks in Russell 3000	1984–1999	The impact of firm-specific characteristics like size and book-to-price on future excess stock returns varies considerably over time. The impact can be either positive or negative at different times. This time variation is partially predictable
Lam (2002)	Global Finance Journal	Stocks listed in the Stock Exchange of Hong Kong	1984–1997	The relation between average returns and size is positive instead of negative. Size, book-to-market equity, and E/P ratios, seem able to capture the cross-sectional variation in average monthly returns. The other two variables, book leverage and market, are also able to capture the cross-sectional variation in average monthly returns. But their effects considered to be redundant when size, book-to-market equity, and E/P ratios are also considered
Lam and Spyrou (2003)	Applied Economics Letters	Nonfinancial listed in the Stock Exchange of Hong Kong	1987–2000	There is a strong positive and significant relationship between size and average returns and that beta alone cannot explain average returns. Average returns and book-to-market have a negative and significant relationship even when size is included in the regression, and there is a negative and significant relationship between returns and ln (asset/market value) and a positive and significant relationship between returns and ln (asset/book value). Finally, E/P does not affect returns and positive cash flow has a positive and significant relationship with returns, even when size is included in the regression
Wang and Xu (2004))	Financial Analysts Journal	A-shares in the Chinese equity market	1996–2002	Beta did not account for return differences among individual stocks. The book-to-market ratio is useless in explaining the cross-sectional differences in equity returns. The size effect, however, continues to be useful, although it is somewhat weaker for Chinese stocks than it has been found to be for US stocks
Wang and Di Iorio (2007)	Global Finance Journal	A-shares in the Chinese equity market	1994–2002	Our findings fail to document an important relationship between beta and stock returns even when the beta effect is examined in up and down-month markets respectively. Fama and French (1992) argue that the correlation coefficient between beta and size is probably −0.98 in the U.S. market. However, the coefficient between beta and size in the A-share market is 0.08 and the performance of the size factor is noteworthy. The value effect is also strong, the positive book-to-market ratio effect was especially evident before June 1997, and the average coefficient of the size effect is not significant in January. There is no factor that has a persistent effect on stock returns, showing that the investment philosophy for Chinese domestic investors changes constantly

performing regressions between each of the candidate factors and stock return to confirm the effect of each factor, and performing regressions between combinations of factors and return to exclude any factor whose effect can be captured by other candidate factors.

Specifically, based on monthly data, the cross-section of stock returns is regressed on explanatory variables as

$$R_{it} = a + b_{1t}V_{1it} + b_{2t}V_{2it} + b_{3t}V_{3it} \ldots + b_{Nt}V_{Nit} + \varepsilon_{it}$$

The subscript t represents period t. R_{it} represents the return from individual stock i from period $t-1$ to t. Then a, b_{1t}, b_{2t}, $b_{3t}\ldots b_{Nt}$ are allowed to vary stochastically in different period. The subscript N represents the number of explanatory variables. Each V represents an explanatory variable; Fama and French (1992) selected effective factors as explanatory variables according to the finding of former researches. And ε_{it} is the residual error of the regression, assumed to have zero mean and to be independent of all other explanatory variables. The regression will be firstly run on each of the explanatory variables to examine the effect of individual factor, and then run on various combinations of factors to clarify the substitutional relations among factors. Finally, the standard t-test was performed to the mean slope coefficients of these regressions (a, b_{1t}, b_{2t}, $b_{3t}\ldots b_{Nt}$) to examine their effects on average returns.

3.2. Causal discovery with ANMCPT

Taking the advantages of data mining techniques in analyzing the big data from stock markets, which are high-dimensional, noisy and dynamic, we applied ANMCPT (Chen et al. 2014) to recognize the directions of many-to-one causalities between portfolio selection factors and stock returns.

3.2.1. ANM

ANM is developed for causal discovery in non-experimental data, and it is effective in discovering one-to-one causalities. It determines the cause and effect in one-to-one causality as follows: if the potential causality can be inferred in one direction but not in the reversed direction, then this direction can be confirmed as causal. More specifically, for two variables x and y, the ANM recognizes the model in one direction,

$$y = f(x) + n, x \perp n.$$

But it does not recognize the reversed model,

$$x = g(y) + n', \; y \perp n'.$$

Here, both n and n′ are additive noise in the above two potential causalities, $x \perp n$ means x and n are both statistically and Gaussian independent on each other. The former causal direction will be inferred as the causality between the factors x and y.

3.2.2. ANMCPT

ANM is an effective one-to-one causal inference method for identifying causalities from synthetic data with additional noise, with a high accuracy of more than 90% (Peters et al. 2011). In stock markets, however, most causal relations are not one-to-one but many-to-one, which ANM cannot distinguish clearly. In this study, the proposed ANMCPT algorithm uses the conditional probability table (CPT) method to transform many-to-one causal inference into one-to-one causal inference in ANM; thus it can be used for many-to-one causal discovery in stock markets. To capture the concept drifts of effective factors, the model is trained by a sliding window method based on rolling years of samples; sliding window is the most popular approach to address concept drift in technical analysis.

The CPT is defined as a set of discrete (not independent) random variables to demonstrate the marginal probability of a single variable with respect to others. For instance, there are m variables $x_1, \cdots, x_m$ and the corresponding numbers of states $K_1, \cdots, K_m$. The CPT of each variable x_i provides the marginal probability value as $p(x_i|x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_m)$, and it has $\prod_{i=1}^{m} K_i$ states.

The discovery process of ANMCPT is composed by the dependence ranking process and the causality discovery process, which are shown in Fig. 1 and Fig. 2.

(1) Dependence ranking process. This process discovers the potential causes and effects of each factor $x_L \in X$ by maximizing the dependence statistics among them. The dependence between factors are tested by Chi-square tests on discrete data. High dependence between two factors means that there exists a potential causality with high probability. The top m factors in dependence ranking are selected as the potential causes/effects, and only the causalities between factor x_L and its potential causes/effects are selected. The dependence ranking process improves the accuracy and effectiveness of ANMCPT for discovering causalities among factors.

(2) Causality discovery process. Each factor $x_L \in X$ has a certain probability of having a single cause or multiple causes. Using these probabilities, ANMCPT discover causalities between each factor x_L and its m potential causes/effects. Because the one-to-one causality is a rare but important causal structure in stock trend prediction, we first infer the causality between factor x_L and each individual factor x_p in its m potential causes/effects forms. When none of the factors x_p can be inferred as the cause of x_L in its potential causes/effects forms, factor x_L may has multiple causes. Then we infer many-to-one causalities by converting the set of factors into a single (dummy) factor with CPT. For instance, the set $S = \{x_{p_1}, \cdots, x_{p_q}\}$ is a subset of size q containing potential causes/effects of factor x_L. The CPT is employed to convert S into single factor x_{SC}, and thus inferring the causality $x_{SC} \rightarrow x_L$ is equal to inferring the causality $S \rightarrow x_L$.

4. Experimental data

In order to verify the effectiveness of ANMCPT, and confirm what types of concept drifts exist in recent China stock market as well as their influences, two datasets were prepared; one is low-dimensional with 8 input variables, the other is high-dimensional with 53 input variables, and both contained one output variable (stock return).

4.1. Dataset I

We collected accounting data and stock prices of all non-financial stocks that have ever appeared in China stock market from Guotaian Finance Database from July 1999 to June 2012. We excluded financial firms because high leverage for non-financial firms is more likely to indicate distress, whereas high leverage is normal for financial firms. Following (Fama and French 1992), we matched the accounting data for all fiscal years ending in calendar year $t-1$ with the returns from July of year t to June of year $t+1$ (t *ranges from 1999 to 2011*). It means that we set a 6-month gap between fiscal yearend and the return tests, which aimed to ensure that the accounting variables were known before the returns were fetched, because listed company were required to publish their annual reports before April 30 whereas some reports may delay. The returns were defined as the first difference of the logarithmic price level. Generally speaking, cross-sectional analysis involves variety of factors that have influence on or reflection to the companies. However, in this study, we just focused on those popular factors adopted in previous cross-sectional researches using Fama–French framework. The selected accounting variables and the corresponding descriptions are listed in Table 2.

4.2. Dataset II

In addition to the factors examined in the present paper, we collected as many other relevant factors as possible from previous

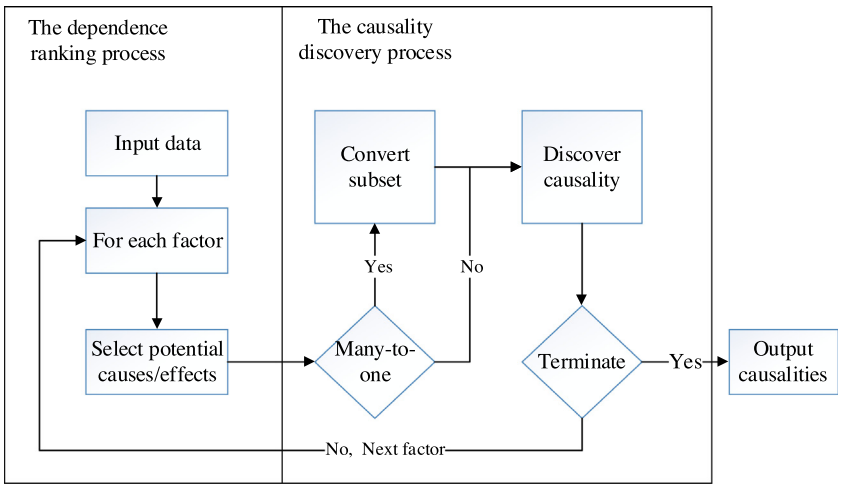

Fig. 1. The framework of ANMCPT.

Line 1: Input data X
Line 2: **For each** factor $x_L \in X$
Line 3: /* the dependence ranking process */
Line 4: Ranking the rest factors $x \in X - \{x_L\}$ by testing the dependence with Chi-square test.
Line 5: The best m factors in dependence ranking as the causes/effects of factor x_L.
Line 6: /* the causality discovery process */
Line 7: Inferring the one-to-one causalities between x_L and $\forall x \in$ causes/effects with ANM.
Line 8: If the cause of label factor x_L have not inferred, then inferring the many-to-one causalities between $S \subset$ causes/effects and x_L with ANMCPT.
Line 9: **End for**
Line 10: Output the cause(s) of each factor $x_L \in X$.

Fig. 2. The algorithm of ANMCPT.

Table 2
Description of Dataset I.

Variables	Description
ln(S)	The log of stock market capitalization
Beta	Market Beta
ln(B/P)	The log of book-to-price ratio
ln(A/P)	Market leverage (the log of total assets/market value)
ln(A/B)	Book leverage(the log of total assets/book value)
E/P	Earning-to-price ratio
C/P	Positive cash flow to price
C/P dummy	Negative cash flow to price

studies on stock prediction (Kiang et al. 1993, Lam 2004, Tsai et al. 2011, Tsai and Hsiao 2010, Zhang et al. 2014). Considering the limited available data, totally 53 factors were selected as listed in Table 3 and definitions of these factors are given in Appendix A. In literature, these factors were used for comprehensive stock evaluation and portfolio selection with regard to various perspectives, including valuation, profitability, growth, leverage, liquidity, operation, momentum, size, cash flow, bonus, volatility, and volume.

5. Experiment results and analyses

5.1. Experiments on cross-section regression on Dataset I

All stocks were sorted in an ascending order and divided into five portfolios by size (*Panel A*), Beta (*Panel B*) and B/P (*Panel C*) respectively, and every portfolio in each panel contained 20 percent of the stock pool. Portfolios are formed yearly. The values of Beta were calculated by regressing the monthly returns of each individual stock on the monthly returns of Shanghai synthesis index. Table 4 reports the average return, beta, size, B/P, leverage, E/P, and cash-flow-to-price of these portfolios.

Panel A shows strong evidence of the size effect. As the firm size grows, the portfolios' average return decreases strictly. The average return on a small-size portfolio, Size 1, is 10.78 percent higher per year than that of a big-size portfolio, Size 5. Our result is consistent with the findings of former researches on Chinese stock market, indicating that size effect is negative (Wang and Xu 2004, Wang and Di Iorio 2007). On the contrary, research results on Hong Kong stock market (Lam and Spyrou 2003) showed that as size increases, beta and return also moves up. In another comparison, as firm size grows, most other accounting ratios listed in *Panel A* (except ln(A/B)) moves up. However, beta has no apparent correlation with firm size. According to (Fama and French 1992), the correlation between beta and size was probably -0.98 in the U.S. market, though the period of data in (Fama and French 1992) war far from ours. The above results indicate that horizontal concept drifts in the correlation between factors may exist among different markets.

Panel B also shows that the Beta has virtually no explanatory value on other variables similar to the result presented in (Wang and Xu 2004, Wang and Di Iorio 2007), but different from the result presented in (Fama and French 1992) that indicated the correlation coefficient between beta and size is probably−0.98.

Moreover, *Panel C* reports the relationship of the growing B/P portfolios with their returns. Findings indicate that the value premium effect is reversed in the China stock market, i.e. the portfolio with a higher B/P has a lower average return, compared to the findings of Fama and French (1998) that value stocks with higher B/P

Table 3
The portfolio selection variables.

Category and features	Category and features
Systematic risk factors	*Operation factors*
(1) Beta	(27) Cash to assets
Valuation factors	(28) Fixed assets turnover
(2) E/P	(29) Inventory turnover
(3) ln(B/P)	(30) Total assets turnover
(4) ln(A/P)	(31) Current assets turnover
(5) ln(A/B)	(32) Long-term liabilities to operating capital
(6) S/P	*Momentum factors*
Profitability factors	(33) Buy–hold return 1-month %
(7) ROE	(34) Buy–hold return 3-month %
(8) ROA	(35) Buy–hold return 6-month %
(9) Gross profit margin	(36) Buy–hold return 12-month %
(10) Net profit margin on sales	Size factors
(11) Operating expense ratio	(37) Circulation market value
(12) Financial expense ratio	(38) Total market value
(13) Earnings before interest and tax to operating income	(39) Circulation market value to total market value
Growth factors	(40) ln(total market value)
(14) Operating profit %	(41) ln(circulation market value)
(15) Gross profit %	(42) ln(total assets)
(16) Net income %	*Bonus factors*
(17) Net asset value per share %	(43) Dividend payout ratio
(18) Total assets %	Cash flow factors
(19) Leverage %	(44) C/P
Leverage factors	*Volatility factors*
(20) Equity to debt ratio	(45) Amplitude 6-month %
(21) Asset liability ratio	(46) Amplitude 12-month %
(22) Long-term liabilities ratio	(47) SD of daily return rate 3-month
(23) Fixed assets ratio	(48) SD of daily return rate 6-month
(24) Current liabilities rate	(49) SD of daily return rate 12-month
Liquidity factors	*Volume factors*
(25) Current ratio	(50) Turnover rate 1-month %
(26) Quick ratio	(51) Turnover rate 3-month %
	(52) Turnover to total market turnover 1-month %
	(53) Turnover to total market turnover 3-month %

tend to have higher profit in United States, Japan, Great Britain, France, Germany, Italy, Netherlands, Belgium, Switzerland, Sweden, Australia, Hong Kong, and Singapore markets. This also means that horizontal concept drifts may exist between China and other mature markets. Wang and Xu (2004) suggested that book-to-market ratio was useless in explaining the cross-sectional differences in equity returns, and Wang and Di Iorio (2007) suggested that the value effect in Chinese stock market was strong and positive especially before June 1997. Both of these results are different from ours. This finding may suggest that vertical concept drifts exist in value effect in Chinese stock market.

Table 5 presents the average of slope coefficients we obtained from 156 monthly regressions of stock returns on size, beta and other explanatory variables that were investigated in (Fama and MacBeth 1973). Standard *t*-tests were performed to the mean slope coefficients of these regressions to confirm their effect on average returns. The effect of ln(B/P) is strong and negative in all of the models. The results further confirm the reversion of value premium in China stock market. In addition, concept drifts were observed with regard to the substitutional relations among factors. Different from (Fama and French 1992), the effect of E/P is no longer absorbed by Size and B/P, whereas the effect of A/P is absorbed by Size. It's worth noting that the strong combinatory effects of A/P and A/B can be explained by the strong effect of B/P in former research as ln(B/P) = ln(A/P) – ln(A/B) (Fama and French 1992). The change of the effect of A/P indicates that the internal logic of B/P effect has changed.

5.2. Experiments on ANMCPT

In accordance with the dividing rules used in Table 4, and meeting the requirement of ANMCPT for discrete variables, the sample in Dataset II were pre-processed as follows:

Table 4
Results for the various portfolios.

Panel A: Size portfolios	Size 1	Size 2	Size 3	Size 4	Size 5
Return (%)	10.83	5.98	3.24	1.41	0.05
Beta	1.03	1.07	1.08	1.06	1.02
ln(S)	6.95	7.47	7.84	8.28	9.26
ln(B/P)	−1.46	−1.15	−0.99	−0.94	−0.78
ln(A/P)	−0.67	−0.34	−0.21	−0.15	0.01
ln(A/B)	0.91	0.83	0.80	0.79	0.79
E/P	−0.01	0.01	0.02	0.04	0.06
C/P	0.48	0.67	0.76	0.87	1.09
C/P Dummy	0.46	0.65	0.72	0.81	0.99
Panel B: Beta portfolios	Beta 1	Beta 2	Beta 3	Beta 4	Beta 5
Return (%)	4.34	5.53	5.53	4.82	1.22
Beta	0.76	1.00	1.08	1.15	1.26
ln(S)	8.03	8.06	7.88	7.88	7.99
ln(B/P)	−1.32	−1.04	−0.99	−0.95	−1.01
ln(A/P)	−0.62	−0.26	−0.18	−0.14	−0.18
ln(A/B)	0.85	0.80	0.82	0.81	0.84
E/P	0.02	0.03	0.02	0.02	0.02
C/P	0.63	0.83	0.84	0.81	0.76
C/P Dummy	0.58	0.76	0.79	0.77	0.72
Panel C: Book-to-market value (B/P) portfolios	B/P 1	B/P 2	B/P 3	B/P 4	B/P 5
Return (%)	7.77	4.97	4.57	2.83	1.56
Beta	0.99	1.06	1.08	1.09	1.04
ln(S)	7.68	7.83	7.99	8.14	8.18
ln(B/P)	−2.32	−1.39	−0.98	−0.60	−0.03
ln(A/P)	−1.21	−0.57	−0.20	0.16	0.45
ln(A/B)	1.12	0.82	0.77	0.75	0.65
E/P	−0.008	0.01	0.02	0.04	0.06
C/P	0.28	0.48	0.68	0.97	1.47
C/P Dummy	0.26	0.45	0.64	0.91	1.37

(1) For each variable, we sorted the sample in ascending order and equal-frequently divided them into five groups (i.e., each group had approximately the same number of sample), and then replaced the values of all samples in each of the five group with 0, 1, 2, 3 and 4, respectively.
(2) For the stock return variable, we used 0 to indicate a negative return and 1 for a positive return.

5.2.1. Re-analyses and results on Dataset I

To verify the effectiveness of ANMCPT, we compared the difference of identified informative factors between cross-section regression approach and ANMCPT using the same dataset. Specifically, we utilized ANMCPT to re-analyze Dataset I (using the entire 13-year data). The three most important causal factors of stock returns were identified as ln(S), ln(B/P) and E/P. This result is consistent with that of the cross-section regression method (see Table 4 and Table 5), thus it provides evidence that the ANMCPT method could be used for knowledge discovery.

5.2.2. Results on Dataset II

To evaluate the ANMCPT's abilities in big data processing and concept drift analysis, we applied ANMCPT to Dataset II. After several trials, we determined to divide and group the 13-year data into 11 subsets, each of which contains a consecutive data of 3 years (such as 1998–2000 and 1999–2001). We adopted this sliding window method (with length of 3) to reduce the negative impact of noise data and thus produce more stable results (compared to the way that data were processed by each year separately).

The result is presented in Table 6. Similar to previous results, the systemic risk factor (#1 Beta) was not identified (as causal factor) by ANMCPT in any period, which implies that Beta is (almost) useless in explaining the cross-sectional differences in stock returns, even when combined with other popular factors. Furthermore, several concept drifts were identified in the relationship among factors and stock returns. Specifically, in the period between 1998 and 2002, size factors (#37 Circulation market value, #38 Total market value, #40 ln(total market value), #41 ln(-circulation market value), #42 ln(total assets)) and volatility factors (#45 Amplitude 6-month%, #46 Amplitude 12-month%, # 47 SD of daily return rate 3-month, #48 SD of daily return rate 6-month, #49 SD of daily return rate 12-month) were informative for investment decision. However, between 2002 and 2005, the informative factors were valuation factors (#2 E/P) and profitability factors (#7 ROE, #8 ROA, #10 Net profit margin on sales, #13 Earnings before interest and tax to operating income). And then during 2005 and 2010, momentum factors (#33 Buy–hold return 1-month%, #34 Buy–hold return 3-month%, #35 Buy–hold return 6-month%, #36 Buy–hold return 12-month%) and volume factors (#52 Turnover to total market turnover 1-month%, #53 Turnover to total market turnover 3-month%) were more important. In addition, B/P seems not so important when compared with the result in Table 5, probably because the effect of B/P is redundant when considering other factors.

To examine the performance of ANMCPT in a data-driven trading system based on data mining techniques, and evaluate the influence of concept drifts in stock prediction, we utilized ANMCPT as a feature selection method (i.e., screening out from a mass of input variables the most informative ones before subsequent model training), and then constructed a variety of prediction models and compared the results. Specifically, we chose five

Table 5
Cross-section regressions.

No.	Constant	Beta	ln(S)	ln(B/P)	ln(A/P)	ln(A/B)	E/P	C/P	C/P dummy
1	1.73 (7.37)*	−0.072 (−0.33)							
2	13.45 (30.82)*		−1.48 (−27.25)*						
3	13.48 (27.62)*	−0.024 (−0.11)	−1.48 (−27.24)*						
4	1.16 (13.31)*			−0.47 (−7.40)*					
5	0.89 (8.05)*				−0.46 (−6.97)*	0.77 (7.29)*			
6	1.74 (30.24)*						−3.61 (−5.60)*		
7	1.67 (25.25)*							0.62 (1.56)	−0.68 (−1.69)
8	13.08 (27.35)*		−1.47 (−26.16)*		0.027 (0.41)	0.44 (4.26)*			
9	13.89 (30.46)*		−1.54(−26.85)*				2.21 (3.32)*		
10	13.87 (31.49)*		−1.55 (−28.00)*					2.29 (5.90)*	−2.21 (−5.55)*
11	1.71 (25.72)*						−4.08 (−6.00)*	1.20 (2.95)*	−1.22 (−2.95)*
12	13.57 (28.43)*		−1.52 (−26.22)*	−0.15 (−2.29)*			2.559 (3.75)*		
13	13.40 (27.56)*		−1.53 (−26.20)*		−0.032 (−0.47)	0.49 (4.70)*	2.46 (3.59)*		
14	14.07 (30.83)*		−1.58 (−27.36)*				1.16 (1.68)	2.15 (5.43)*	−2.08 (−5.14)*

Notes: The average slope is the time-series average of the monthly regression slopes. The *t*-statistics appear in parentheses under the slopes.
* Denotes significance at the 5% level.

Table 6
Result of ANMCPT in Dataset II.

Year/factors #.	2	3	4	7	8	10	13	30	33	34	35	36	37	38	40	41	42	45	46	47	48	49	52	53
1998–2000		*											*			*			*					*
1999–2001																		*	*		*	*		*
2000–2002														*	*		*		*	*				
2001–2003					*						*	*		*								*		
2002–2004	*			*	*									*								*		
2003–2005	*			*	*	*	*																	
2004–2006	*											*							*		*	*		
2005–2007			*						*	*	*	*												
2006–2008									*	*	*			*									*	
2007–2009									*	*	*			*									*	
2008–2010								*	*		*												*	*

Note: Only variables which are selected by ANMCPT at least once would be presented in the table. Factors are highlighted and grouped according to Table 3: factors 2, 3 and 4 belong to valuation factors; factors 7, 8, 10 and 13 belong to profitability factors; factor 30 belongs to operation factors; factors 33, 34, 35 and 36 belong to momentum factors; factors 37, 38, 40, 41 and 42 belong to size factors; factors 45, 46, 47, 48 and 49 belong to volatility factors; and factors 52 and 53 belong to volume factors.

Table 7
The prediction accuracy of five popular algorithms.

Algorithm/years	1998–2000	1999–2001	2000–2002	2001–2003	2002–2004	2003–2005	2004–2006	2005–2007	2006–2008	2007–2009	2008–2010
ANMCPT-based models											
ANMCPT + Logistic	56.06	72.01	88.85	88.83	59.16	69.12	70.28	65.97	60.39	71.76	56.17
ANMCPT + J48	60.85	72.01	88.85	88.83	58.29	69.12	70.28	66.22	60.5	70.61	55.27
ANMCPT + NB	56.62	67.58	83.16	82.59	57.65	65.56	69.12	64.89	60.82	71.01	56.6
ANMCPT + BAN	56.49	67.58	83.12	82.48	57.62	65.53	69.12	64.89	60.82	71.04	56.6
ANMCPT + TAN	61.09	71.1	88.85	88.79	57.38	68.78	70.19	65.86	60.39	71.73	55.7
NoFS (53)											
Logistic	48.42	67.82	87.08	87.42	51.41	68.01	68.51	64.2	59.15	70.13	53.82
J48	62.52	72.16	88.85	88.83	58.89	67.36	69.52	65.95	62.87	69.78	56.01
NB	56.56	64.74	77.9	75.58	53.32	60.03	61.56	62.21	61.03	68.1	56.01
BAN	56.56	64.74	77.9	75.5	53.32	60	61.62	62.21	61.03	68.07	56.01
TAN	64.51	70.38	85.06	84	63.05	71.31	71.3	66.33	63.35	70.02	58.96
NoFS (8)											
Logistic	52.13	71.61	88.67	88.66	57.58	69.34	70.55	66.34	57.93	71.03	55.76
J48	53.57	71.90	88.67	88.66	56.80	69.37	70.55	66.47	58.43	71.26	54.82
NB	52.13	71.46	88.03	88.18	56.12	69.37	70.55	64.80	56.74	69.48	56.02
BAN	52.13	71.46	88.12	88.18	56.12	69.37	70.55	64.77	56.74	69.48	56.02
TAN	52.19	69.42	87.61	87.65	56.12	68.00	69.93	64.69	57.08	70.42	54.44
ANMCPT – NoFS (53)											
Logistic	7.64	4.19	1.77	1.41	7.75	1.11	1.77	1.77	1.24	1.63	2.36
J48	−1.68	−0.14	0	0	−0.6	1.76	0.76	0.28	−2.37	0.83	−0.74
NB	0.06	2.84	5.26	7.02	4.33	5.53	7.55	2.69	−0.22	2.91	0.58
BAN	−0.06	2.84	5.22	6.98	4.3	5.53	7.5	2.69	−0.22	2.97	0.58
TAN	−3.42	0.72	3.79	4.79	−5.67	−2.53	−1.1	−0.47	−2.96	1.71	−3.26
ANMCPT – NoFS (8)											
Logistic	3.93	0.4	0.18	0.17	1.58	−0.22	−0.27	−0.37	2.46	0.73	0.41
J48	7.28	0.11	0.18	0.17	1.49	−0.25	−0.27	−0.25	2.07	−0.65	0.45
NB	4.49	−3.88	−4.87	−5.59	1.53	−3.81	−1.43	0.09	4.08	1.53	0.58
BAN	4.36	−3.88	−5	−5.7	1.5	−3.84	−1.43	0.12	4.08	1.56	0.58
TAN	8.9	1.68	1.24	1.14	1.26	0.78	0.26	1.17	3.31	1.31	1.26
ANMCPT > NoFS (53)	2	4	5	5	3	4	4	4	1	5	3
ANMCPT > NoFS (8)	5	3	3	3	5	1	1	3	5	4	5

Note: The *ANMCPT* panel presents the prediction accuracy obtained when ANMCPT was used for feature selction. The *noFS* (8) panel presents the obtained prediction accuracy without factors selection in dataset I and *noFS* (53) presents the obtained prediction accuracy without factors selection in dataset II. The *ANMCPT-NoFS* panel presents the difference of prediction accuracy between *ANMCPT* panel and *NoFS panel*.

popular data mining algorithms implemented in a famous data mining software–Weka (http://www.cs.waikato.ac.nz/ml/weka/), including Logistic regression, J48 (a classical decision tree algorithm), Naïve Bayes (NB), BN Augmented Naïve Bayes (BAN) and Tree Augmented Naïve Bayes (TAN).

Table 7 shows the prediction accuracy results. Each prediction accuracy was obtained using 10-fold cross-validation based on the corresponding data subset. The '*ANMCPT-NoFS (53)*' panel shows that the ANMCPT can improve prediction performance in 39 out of the total 55 tests (73%). Moreover, the '*ANMCPT*' panel contains two periods (1999–2001 and 2000–2002) with prediction accuracy higher than 80% in all five algorithms, and no period in '*NoFS (53)*' panel can reach this extent. These results indicate that ANMCPT is effective for feature selection. Meanwhile, '*ANMCPT-NoFS (8)*' shows that using the factors selected by ANMCPT from dataset II is better than using the factors considered in most former cross-sectional researches to construct prediction models; ANMCPT outperforms in 9 of the total 11 periods (80%) and 38 of the total 55 tests (nearly 70%). These results demonstrate the effectiveness of ANMCPT and the significant influence of concept drifts on stock prediction.

6. Conclusions

In conventional cross-sectional analysis, the regression approach proposed by Fama and French (1992) is used to investigate the relationships among explanatory variables and stock returns. However, this approach requires expert knowledge, which has obvious disadvantages when handling problems that involve concept drifts and high-dimensional data. Therefore, in this study, we applied ANMCPT, which is derived from a popular observational data-based causal inference method called ANM, to identify causal relationships in China stock market (covering a long period from July 1999 to June 2012).

In our experiments on the low-dimensional dataset of 8 factors (Dataset I), firstly, we conducted multivariate cross-sectional analysis, and found that: (1) the reversion of value premium (B/P effect) is confirmed; (2) as the explanatory power of A/P is absorbed by Size, the internal cause of B/P effect has changed; (3) E/P effect can no longer be captured by B/P and size, and size effect seems contrary to the Hong Kong market in a long-term period; and (4) the above results suggest that both vertical (between different markets) and horizontal (between different periods) concept drifts have occurred, and the drifts exist in both the relations between factors and returns and between factors. Then, we applied ANMCPT to Dataset I and compared its results with that of cross-sectional regressions. The results show that ANMCPT can identify similar relationships as the regression approach does, e.g., the significant size and B/P effects.

In our experiments on the high-dimensional dataset of 53 factors (Dataset II), we (can only) applied ANMCPT. Results show that factors selected by ANMCPT outperformed the factors considered in most former cross-sectional researches in more than 80% of the total 11 periods and nearly 70% of the total 55 tests. Moreover, the dynamic nature of the explanatory powers of different factors on future stock returns has been exhibited. For example,

the effect of B/P is weak or redundant when more factors are considered in recent China market, the size effect is significant only between 2006 and 2011, and other informative factors for portfolio selection also change over time.

As the knowledge discovered from historical data might become useless or reversed on new data with unexpected patterns, it is necessary to continuously investigate concept drifts. For instance, it is difficult to determine whether the factors not selected by ANMCPT would always be useless or just useless in some specific testing periods. Moreover, the difference of B/P effect in Dataset I and Dataset II necessitates further consideration of the interactions among factors.

To the best of our knowledge, this paper is the first to empirically compare the conventional cross-sectional regression approach with a causal inference method, and it may motivate a series of future researches towards the application of advanced computational techniques, such as data mining and causal discovery, to address the concept drift issue in cross-sectional analysis of stock markets.

For further studies, we suggest researchers to (1) utilize more advanced causal inference methods to explore continuous stock market data because the discretization of a continuous variable would inevitably lead to the loss of information quality; (2) predict/estimate when a concept drift would occur before it really occurs, rather than merely confirm the previous ones; and (3) investigate the concept drift phenomenon in other markets and other time periods.

Acknowledgments

This research was partly supported by the National Natural Science Foundation of China (71271061, 70801020), Science and Technology Planning Project of Guangdong Province, China (2010B010600034, 2012B091100192), Guangdong Natural Science Foundation Research Team (S2013030015737), and Business Intelligence Key Team of Guangdong University of Foreign Studies (TD1202). The authors are grateful for the constructive comments of the two referees on the earlier versions of this paper.

Appendix A. Factor definitions

(#) Factor	Definition
Beta	The market beta is estimated based on returns of twenty-four months
E/P	Earnings-to-price ratio = Earnings per share/price per share
ln(B/P)	The natural logarithm of Book-to-price ratio = ln(Book value/price per share)
ln(A/P)	The natural logarithm of Market leverage = ln(total assets/market value)
ln(A/B)	The natural logarithm of Book leverage = ln(total assets/book value)
S/P	Sales-to-price ratio = Sales/price per share
ROE	Return on equity = net profit after tax before extraordinary items/shareholders' equity
ROA	Return on assets = net profit after tax before extraordinary items/total assets
Gross profit margin	(Net sales–cost of goods sold)/net sales
Net profit margin on sales	Net income after tax/net sales
Operating expense ratio	Operating expense/operating income
Financial expense ratio	Financial expense/operating income
Earnings before interest and tax to operating income	Earnings before interest and tax/operating income
Operating profit%	Operating profit growth rate = (operating profit at the current year − operating profit at the previous year)/operating profit at the previous year
Gross profit%	Gross profit growth rate = (gross profit at the current year − gross profit at the previous year)/gross profit at the previous year
Net income%	Net income growth rate = (net income after tax at the current year − net income after tax at the previous year)/net income after tax at the previous year
Net asset value per share%	Net asset value per share growth rate = (net asset value per share at the current year− net asset value per share at the previous year)/net asset value per share at the previous year
Total assets%	Total assets growth rate = (total assets at the current year − total assets at the previous year)/total assets at the previous year
Leverage%	Debt to equity growth rate = (debt to equity at the current year − debt to equity at the previous year)/debt to equity at the previous year
Equity to debt ratio	Shareholders' equity/total liabilities
Asset liability ratio	Total assets/total liabilities
Long-term liabilities ratio	Long-term liabilities/total liabilities
Fixed assets ratio	Net fixed assets/total assets
Current liabilities rate	Current liabilities/total liabilities
Current ratio	Current assets/current liabilities
Quick ratio	(current assets − inventory)/current liabilities
Cash to assets	Operating net cash flow/total assets
Fixed assets turnover	Operating income/average fixed assets, where average fixed assets = (fixed assets closing balance + fixed assets opening balance)/2
Inventory turnover	Operating cost/average inventory, where average inventory = (inventory closing balance + inventory opening balance)/2

(*continued on next page*)

Total assets turnover	Operating income/total average assets, where total average assets = (total assets closing balance + total assets opening balance)/2
Current assets turnover	Operating income/average current assets, where average current assets = (current assets closing balance + current assets opening balance)/2
Long-term liabilities to operating capital	Long-term liabilities/operating capital
Buy–hold return 1-month%	(Closing price of the last trading day – closing price of the first trading day)/closing price of the first trading day * 100% (in the recent month)
Buy–hold return 3-month%	(Closing price of the last trading day – closing price of the first trading day)/closing price of the first trading day * 100% (in the recent 3 months)
Buy–hold return 6-month%	(closing price of the last trading day – closing price of the first trading day)/closing price of the first trading day * 100% (in the recent 6 months)
Buy–hold return 12-month%	(closing price of the last trading day – closing price of the first trading day)/closing price of the first trading day * 100% (in the recent 12 months)
Circulation market value	Price per share * number of shares outstanding
Total market value	Price per share * capital size
Circulation market value to total market value	Circulation market value/total market value
ln(Size)	The natural logarithm of total market value
ln(circulation market value)	The natural logarithm of circulation market value
ln(total assets)	The natural logarithm of total assets
C/P	(Operating cash inflows – operating cash outflows per share)/price per share
Dividend payout ratio	Dividend per share/earnings per share
Amplitude 6-month%	(the highest price of the current 6 months − the lowest price of the current 6 months)/the closing price of the previous 6 months
Amplitude 12-month%	(the highest price of the current 12 months − the lowest price of the current 12 months)/the closing price of the previous 12 months
SD of daily return rate 3-month	The deviation amplitude of the daily return rate relative to average daily return rate in 3 months
SD of daily return rate 6-month	The deviation amplitude of the daily return rate relative to average daily return rate in 6 months
SD of daily return rate 12-month	The deviation amplitude of the daily return rate relative to average daily return rate in 12 months
Turnover rate 1-month%	Trading volume/number of shares outstanding (in 1 month)
Turnover rate 3-month%	Trading volume/number of shares outstanding (in 3 months)
Turnover to total market turnover 1-month%	Turnover/total market turnover (in 1 month)
Turnover to total market turnover 3-month%	Turnover/total market turnover (in 3 months)

References

Allen, F., Karjalainen, R., 1999. Using genetic algorithms to find technical trading rules. Journal of Financial Economics 51 (2), 245–271.

Atsalakis, G.S., Valavanis, K.P., 2009. Surveying stock market forecasting techniques–Part II: soft computing methods. Expert Systems with Applications 36 (3), 5932–5941.

Bahrammirzaee, A., 2010. A comparative survey of artificial intelligence applications in finance: artificial neural networks, expert system and hybrid intelligent systems. Neural Computing and Applications 19 (8), 1165–1195.

Ball, R., 1978. Anomalies in relationships between securities' yields and yield-surrogates. Journal of Financial Economics 6 (2), 103–126.

Banz, R.W., 1981. The relationship between return and market value of common stocks. Journal of Financial Economics 9 (1), 3–18.

Basu, S., 1983. The relationship between earnings' yield, market value and return for NYSE common stocks: further evidence. Journal of Financial Economics 12 (1), 129–156.

Bhandari, L.C., 1988. Debt/equity ratio and expected common stock returns: Empirical evidence. The Journal of Finance 43 (2), 507–528.

Bilson, C.M., Brailsford, T.J., Hooper, V.J., 2001. Selecting macroeconomic variables as explanatory factors of emerging stock market returns. Pacific-Basin Finance Journal 9 (4), 401–426.

Black, F., 1972. Capital market equilibrium with restricted borrowing. Journal of Business, 444–455.

Bodie, Z., Kane, A., Marcus, A., 2005. Investments, 6th edition. Mcgraw-Hill, New York.

Chan, L.K., Karceski, J., Lakonishok, J., 2000. New paradigm or same old hype in equity investing? Financial Analysts Journal 56 (4), 23–36.

Chen, W., Liu, K., Su, L., Liu, M., Hao, Z., Hu, Y., and Zhang, X. Discovering many-to-one causality in Software Project Risk Analysis. Paper presented at the 9TH International Conference on P2P, Parallel, Grid, Cloud and Internet Computing, Guangzhou, China, 2014.

Delany, S.J., Cunningham, P., Tsymbal, A., Coyle, L., 2005. A case-based technique for tracking concept drift in spam filtering. Knowledge-Based Systems 18 (4), 187–195.

Fama, E.F., French, K.R., 1992. The cross-section of expected stock returns. The Journal of Finance 47 (2), 427–465.

Fama, E.F., French, K.R., 1996. Multifactor explanations of asset pricing anomalies. The Journal of Finance 51 (1), 55–84.

Fama, E.F., French, K.R., 1998. Value versus growth: the international evidence. The Journal of Finance 53 (6), 1975–1999.

Fama, E.F., MacBeth, J.D., 1973. Risk, return, and equilibrium: empirical tests. The Journal of Political Economy, 607–636.

Filis, G., 2010. Macro economy, stock market and oil prices: do meaningful relationships exist among their cyclical fluctuations? Energy Economics 32 (4), 877–886.

Garcia, M.E.F., Enrique, A., Garcia, R.Q., 2010. Improving return using risk-return adjustment and incremental training in technical trading rules with GAPs. Applied Intelligence 33 (2), 93–106.

Gay Jr., R.D., 2008. Effect of macroeconomic variables on stock market returns for four emerging economies: Brazil, Russia, India, and China. International Business and Economics Research Journal 7 (3), 1–8.

Gorgulho, A., Neves, R., Horta, N., 2011. Applying a GA kernel on optimizing technical analysis rules for stock picking and portfolio composition. Expert Systems with Applications 38 (11), 14072–14085.

Hadavandi, E., Shavandi, H., Ghanbari, A., 2010. Integration of genetic fuzzy systems and artificial neural networks for stock price forecasting. Knowledge-Based Systems 23 (8), 800–808.

Hendershott, T., Jones, C.M., Menkveld, A.J., 2011. Does algorithmic trading improve liquidity? The Journal of Finance 66 (1), 1–33.
Henry, P.B., 2000. Stock market liberalization, economic reform, and emerging market equity prices. The Journal of Finance 55 (2), 529–564.
Hoyer, P. O., Janzing, D., Mooij, J. M., Peters, J., and Schölkopf, B. Nonlinear causal discovery with additive noise models. Paper presented at the NIPS, 2008.
Hsu, C., 2011. A hybrid procedure for stock price prediction by integrating self-organizing map and genetic programming. Expert Systems with Applications 38 (11), 14026–14036.
Hu, Y., Zhang, X., Ngai, E.W.T., Cai, R., Liu, M., 2013. Software project risk analysis using Bayesian networks with causality constraints. Decision Support Systems 56, 439–449. http://dx.doi.org/10.1016/j.dss.2012.11.001.
Janzing, D., Mooij, J., Zhang, K., Lemeire, J., Zscheischler, J., Daniušis, P., Schölkopf, B., 2012. Information-geometric approach to inferring causal directions. Artificial Intelligence 182, 1–31.
Kaucic, M., 2010. Investment using evolutionary learning methods and technical rules. European Journal of Operational Research 207 (3), 1717–1727.
Kiang, M.Y., Chi, R.T., Tam, K.Y., 1993. DKAS: a distributed knowledge acquisition system in a DSS. Journal of Management Information Systems 9 (4), 59–82.
Kleinberg, S., Hripcsak, G., 2011. A review of causal inference for biomedical informatics. Journal of Biomedical Informatics 44 (6), 1102–1112.
Lam, H.Y., Spyrou, S.I., 2003. Fundamental variables and the cross-section of expected stock returns: the case of Hong Kong. Applied Economics Letters 10 (5), 307–310.
Lam, K.S., 2002. The relationship between size, book-to-market equity ratio, earnings–price ratio, and return for the Hong Kong stock market. Global Finance Journal 13 (2), 163–179.
Lam, M., 2004. Neural network techniques for financial performance prediction: integrating fundamental and technical analysis. Decision Support Systems 37 (4), 567–581.
Lintner, J., 1965. The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. The Review of Economics and Statistics, 13–37.
Liu, C., Yeh, C., Lee, S., 2012. Application of type-2 neuro-fuzzy modeling in stock price prediction. Applied Soft Computing 12 (4), 1348–1358.
Liu, M., Cai, R., Hu, Y., Matheny, M.E., Sun, J., Hu, J., Xu, H., 2014. Determining molecular predictors of adverse drug reactions with causality analysis based on structure learning. Journal of the American Medical Informatics Association 21 (2), 245–251.
Lucas, A., van Dijk, R., Kloek, T., 2002. Stock selection, style rotation, and risk. Journal of Empirical Finance 9 (1), 1–34.
Malkiel, B.G., Fama, E.F., 1970. Efficient capital markets: a review of theory and empirical work*. The Journal of Finance 25 (2), 383–417.
Mani, S., and Cooper, G. F. Causal discovery from medical textual data. Paper presented at the Proceedings of the AMIA Symposium, 2000.
Mani, S., Cooper, G.F., 2004. Causal discovery using a Bayesian local causal discovery algorithm. Medinfo 11 (Pt 1), 731–735.
Matilla-García, M., 2006. Are trading rules based on genetic algorithms profitable? Applied Economics Letters 13 (2), 123–126.
Morck, R., Yeung, B., Yu, W., 2000. The information content of stock markets: why do emerging markets have synchronous stock price movements? Journal of Financial Economics 58 (1), 215–260.
Neely, C.J., 2003. Risk-adjusted, ex ante, optimal technical trading rules in equity markets. International Review of Economics and Finance 12 (1), 69–87.
Pearl, J., 2000. Causality: Models, Reasoning, and Inference. Cambridge University Press, Los Angeles, USA.
Pearl, J., Verma, T.S., 1995. A theory of inferred causation. Studies in Logic and the Foundations of Mathematics 134, 789–811.
Peters, J., Janzing, D., Scholkopf, B., 2011. Causal inference on discrete data using additive noise models. IEEE Transactions on Pattern Analysis and Machine Intelligence 33 (12), 2436–2450.
Peters, J., Mooij, J.M., Janzing, D., Schölkopf, B., 2014. Causal discovery with continuous additive noise models. The Journal of Machine Learning Research 15 (1), 2009–2053.
Rosenberg, B., Reid, K., Lanstein, R., 1985. Persuasive evidence of market inefficiency. The Journal of Portfolio Management 11 (3), 9–16.
Shao, J., Ahmadi, Z., and Kramer, S. Prototype-based learning on concept-drifting data streams. Paper presented at the Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, 2014.
Sharpe, W.F., 1964. Capital asset prices: a theory of market equilibrium under conditions of risk*. The Journal of Finance 19 (3), 425–442.
Stattman, D., 1980. Book values and stock returns. The Chicago MBA: A journal of selected papers 4 (1), 25–45.
The New York Times. 2012. Times Topics: High-Frequency Trading. Retrieved from <http://topics.nytimes.com/top/reference/timestopics/subjects/h/high_frequency_algorithmic_trading/index.html>.
Tsai, C., Lin, Y., Yen, D.C., Chen, Y., 2011. Predicting stock returns by classifier ensembles. Applied Soft Computing 11 (2), 2452–2459.
Tsai, C., Hsiao, Y., 2010. Combining multiple feature selection methods for stock prediction: union, intersection, and multi-intersection approaches. Decision Support Systems 50 (1), 258–269.
Tsymbal, A. The problem of concept drift: definitions and related work. Computer Science Department, Trinity College Dublin, 2004.
Wang, F., Xu, Y., 2004. What determines Chinese stock returns? Financial Analysts Journal, 65–77.
Wang, H., Fan, W., Yu, P. S., and Han, J. Mining concept-drifting data streams using ensemble classifiers. Paper presented at the Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, 2003.
Wang, Y., Di Iorio, A., 2007. The cross section of expected stock returns in the Chinese A-share market. Global Finance Journal 17 (3), 335–349.
World Federation of Exchanges. 2013. World Federation of Exchanges 2012 Annual Report & Statistics 2013/10/20, from <http://www.world-exchanges.org/files/statistics/pdf/WFE2012%20final.pdf>.
Zhang, X., Hu, Y., Xie, K., Wang, S., Ngai, E.W.T., Liu, M., 2014. A causal feature selection algorithm for stock prediction modeling. Neurocomputing 142, 48–59. http://dx.doi.org/10.1016/j.neucom.2014.01.057.
Zliobaite, I. Learning under concept drift: an overview. Technical report. 2009, Faculty of Mathematics and Informatics, Vilnius University: Vilnius, Lithuania, 2009.