Aggregating multiple types of complex data in stock market prediction: A model-independent framework

Huiwen Wang[a,b], Shan Lu[a,c], Jichang Zhao[a,b,*]

[a]*School of Economics and Management, Beihang University*
[b]*Beijing Advanced Innovation Center for Big Data and Brain Computing*
[c]*Beijing Key Laboratory of Emergency Support Simulation Technologies for City Operations*

Abstract

The increasing richness in volume, and especially types of data in the financial domain provides unprecedented opportunities to understand the stock market more comprehensively and makes the price prediction more accurate than before. However, they also bring challenges to classic statistic approaches since those models might be constrained to a certain type of data. Aiming at aggregating differently sourced information and offering type-free capability to existing models, a framework for predicting stock market of scenarios with mixed data, including scalar data, compositional data (pie-like) and functional data (curve-like), is established. The presented framework is model-independent, as it serves like an interface to multiple types of data and can be combined with various prediction models. And it is proved to be effective through numerical simulations. Regarding to price prediction, we incorporate the trading volume (scalar data), intraday return series (functional data), and investors' emotions from social media (compositional data) through the framework to competently forecast whether the market goes up or down at opening in the next day. The strong explanatory power of the framework is further demonstrated. Specifically, it is found that the intraday returns impact the following opening prices differently between bearish market and bullish market. And it is not at the beginning of the bearish market but the subsequent period in which the investors' "fear" comes to be indicative. The framework would help extend existing prediction models easily to scenarios with multiple types of data and shed light on a more systemic understanding of the stock market.

Keywords: Stock market prediction, Investors' emotions, Complex data, Data aggregation, Model-independent

*Corresponding author: jichang@buaa.edu.cn

1. Introduction

Predicting stock prices have attracted significant research interests in both theories and applications. With the development of technology, the data related to stock market has been increasingly accumulated and diversified in either the sources or the types. For example, the direct sources originate in the financial system itself, such as the price information at various frequencies (Harris, 1986; Jain and Joh, 1988; Pan et al., 2017), the companies financial reports (Jones and Litzenberger, 1970; Zhou et al., 2015, 2017a), and financial news (Geva and Zahavi, 2014; Li et al., 2014; Hagenau et al., 2013). The indirect sources are those outside the financial system, like the rise and fall of macro economic (Chen et al., 1986), the reactions and reflections from investors' emotion revealed by social media (Zhou et al., 2017b; Sun et al., 2017; Ruan et al., 2018; Zhang et al., 2017; Li et al., 2014), search engine (Preis et al., 2013) or analyst's recommendations (Duan et al., 2013), etc. The richness of data sources has provided the chances to understand the stock market more comprehensively and make the price prediction more accurate than before. In the meantime, it brings challenges to the classic statistic analysis since they may not be suitable for dealing with these data at the same time. For example, the investors' emotion data are pie-like whose components are the proportions of different emotions, however, the intraday return series are curve-like. The former type of data is usually considered as compositional data and the latter one is functional data. As the two types of data belong to different spaces, it is not reasonable to directly combine them together and deploy statistic analysis.

The complex data analysis has been rapidly developed in the past decades. Two of the most popular types of complex data are compositional data and functional data. One observation of the compositional data is constituted by D proportions that subject to a unit sum constraint. Since 1896, compositional data has been the focus of research (Pearson, 1896; Chayes, 1960), and has been applied in many research fields such as economics (Longford and Pittau, 2006), ecology (Aebischer et al., 1993; Bingham et al., 2007), geochemistry (Buc-

cianti et al., 2006; Miesch and Chapman, 1977), social science (Godichon-Baggioni et al., 2018), etc. Meanwhile, the studies on functional data analysis (FDA) has grown rapidly (Ramsay and Silverman, 1997; Ferraty and Vieu, 2006; Horváth and Kokoszka, 2012; Fan et al., 2015) in the past decades. One observation of the functional data consists of a function (often a smooth curve, but not always). Functional linear model is among the most popular methods that have been widely used in the FDA (Ramsay and Silverman, 2007; Horváth and Kokoszka, 2012; Cai et al., 2006). Many results have been published on the functional linear model, in which only functional predictor is presented (Hall and Hooker, 2016; Comte et al., 2012; García-Portugués et al., 2014; Escabias et al., 2004; Shang et al., 2015; Huang et al., 2016). Nevertheless, these regression models are constrained to only single type of data and there are few efforts that consider the mixed-type of data when modelling (Wang et al., 2016). And in fact, aggregating data of various sources or formats will indeed enrich the views and resolutions of the explorations.

Stock prices prediction is one of such cases. While the daily price series are the most common data when conducting prediction (Hsu, 2011; Jasemi et al., 2011; Efendi et al., 2018; Baralis et al., 2017; Ye et al., 2016; Chen and Chen, 2015), other kinds of data are attracting people's attention as rapid growth of the generation and reservation of financial data. Firstly, public online emotion from social media has been used in predicting the stock market. Based on sixty thousands of microblogs from Sina Weibo, the largest online social media in China, Wanyun and Jie (2013) demonstrates that the public online emotion could only predict the trading volume rather than the prices. However, since they conduct the study using neither a huge amount of microblogs nor an effective classifier, their results may not be a generic conclusion. Zhou et al. (2017b) assigns five labels, including "anger", "disgust", "joy", "sadness", "fear", to over 3.5 million microblogs and shows that "disgust", "joy", "sadness", "fear" could be useful in predicting the Chinese stock market index. The daily emotions of the five types are naturally compositional data observations. Secondly, as the time series of

intraday prices at various frequency became available, researchers have documented many intraday phenomena that related to stock returns, including the prices rise at the end of the day (Harris, 1986, 1989), significant weekday differences in intraday returns accrue during the first 45 minutes after the market opens (Harris, 1986), largest stock returns occur during the first (except on Monday) and the last trading hours, the lowest average return is earned in the fifth hour of the day (Jain and Joh, 1988). However, to our best knowledge, the exact knowledge on how the intraday returns would influence the future price is still to be discussed except one existing study that endorses autoregressive, random walk linear models, smooth transition, Markov switching, artificial neural network, non-parametric kernel regression and support vector machine models to predict the intraday returns (Matías and Reboredo, 2012). Considering the inconsistent frequencies of the intraday return series with the commonly used daily return series, we argue that the intraday return curves can be used as observations of functional data, that is, one curve for one trading day. Thirdly, daily trading volume has been proved to be an important indicator in stock analysis as it is used to measure the relative worth of a market move (Foster and Viswanathan, 1993; Lillo et al., 2003), and it usually belongs the common scalar data. Given the abundant information of stock market, however, how to integrate the multiple types of data into one prediction model is still unknown but of great importance for understanding and predicting the stock market.

To fill this vital gap, in this study, we propose a framework that incorporates the investors' emotions from social media (compositional data), intraday return series (functional data), and trading volume (scalar data) together to predict whether the market goes up or down at opening in the next day. As the goal of prediction is binary, it could be viewed as a classification problem. Specifically, by transforming the original data in terms of isometric logratio transformation, and functional principal component basis expansion, respectively, we can sufficiently obtain consistent numeric types of features from both compositional data and functional data. Since the transformation is independent to the prediction classifier,

4

the framework serves as an interface between the data and the prediction model. We adopt logistic regression as a case of the classification models and present the corresponding estimation procedure. Note that other classification models could also be combined into the present approach, while logistic regression is particularly useful when the class is dichotomous and it does not need to assume data distributions on variables. More importantly, unlike the "black-box" approaches such as support vector machine, logistic regression could provide the predictors' coefficients, which is important for the model to give more insights on the relationships between the predictors and the response. Due to these benefits, we mainly consider logistic regression in this paper, but the framework can be combined with any prediction model. The estimation procedure of the framework is further proved to be consistent and effective through numerical simulations.

In the real-world application, by dividing the sample period into three phases, the model exhibits a good prediction power when conducting out-of-sample predictions, especially in the first two phases. Besides, we find that both functional coefficients and numeric coefficients shed light on the different market status. Most surprisingly, we find that in the bullish market (phase 1), the "sadness" is more indicative than "joy". And in the initial market crash (phase 2), the "disgust" plays a dominant role in explaining the market. When the market became depressed (phase 3), the "anger" and "fear" begun to do their parts as well as other emotions. Furthermore, our results show that it is not at the beginning of the bearish market but the subsequent period in which the investors' "fear" comes to be indicative.

The rest of the paper is organized as follows. Section 2 introduces the data of the three types of predictors and the binary response that have motivated us to develop the framework with multiple types of data. In section 3, we illustrate the transformation approaches to deal with compositional data and functional data, and then propose the model-independent framework. We also present how to estimate parameters of the logistic regression under the

framework, which is considered as a case of the classification methods. Section 4 performs the simulation studies to prove that the proposed framework could yield effective estimation results. Section 5 presents the results of the prediction on Chinese stock market index, from the view of both explanation and prediction power. In section 6, we draw the conclusions as well as some of the limitations of this paper.

2. Data

2.1. Sample period and binary response

In this study we consider the Chinese stock market, one of the largest markets in the world based on market capitalization. The sample period of this study is 2014/12/02 to 2016/4/29 (345 trading days in total), covering the recent boom and bust of Chinese stock market. As can be seen in Fig. 1, Shanghai Stock Exchange Composite index, which is one of the most important stock market indices in China, kept rising from the end of 2014 to the top of the past seven years at June 2015, and then went down sharply in the following months. From then on, the market kept vibrating at the low level that is close to the end of 2014.

Since the underlying market fundamentals vary a lot across the whole period, we cut the sample period into three phases illustrated by different backgrounds in Fig. 1. The first phase starts from 2014/12/2 to 2015/6/18, which had witnessed the enormous booming of the Chinese stock market. The second phase ranges from 2015/6/19 to 2015/10/14, starting with the popping of the stock market bubble and followed by severe turbulence though government had implemented a lot of bailout measures. The third phase defined in this paper starts from 2015/10/15 to 2016/4/29, when the market suffered from the major systemic aftershock and kept being depressed. Moreover, as will be discussed in section 2.4, the pattern of the market emotions varies with the three phases. Therefore, the following analysis and modelling are applied to the three phases separately.
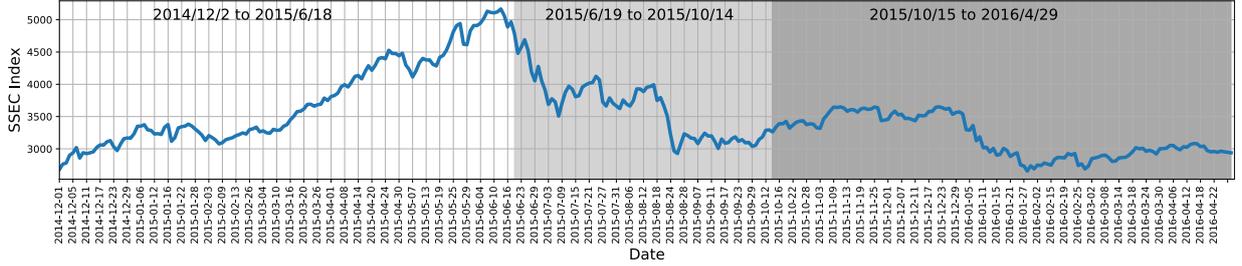
6

Figure 1: **Shanghai Stock Exchange Component index.** The sample period of this study starts from 2014/12/02 to 2016/4/29 (345 trading days in total), and completely covers the recent boom and bust of the Chinese stock market. The sample period is divided into three phases illustrated by different backgrounds.

In this paper, the Shanghai Stock Exchange Composite (SSEC) index is employed as an indicator to represent the trend of Chinese stock market. The closing price of SSEC on day $i - 1$ is denoted as $closing_{i-1}$ and the opening price of SSEC on day i is denoted as $opening_i$. Then the daily open return of SSEC on day i is defined as $y_i^* = (opening_i - closing_{i-1})/closing_{i-1}$. The reason is that this kind of percentage change is consistent with what investors see at the moment of the market opening on any trading information board (Lu et al., 2017). Instead of the exact value of the open return, whether the open return is positive or negative is of the foremost interest in reality, because it could provide advice on the direction of trading. Therefore, using zero as the cut point, the y_i^* is transformed into a binary variable y_i, i.e.

$$
y_i = \begin{cases} 1, & y_i^* > 0 \\ 0, & otherwise. \end{cases} \tag{1}
$$

$y_i, i = 1, ..., n$ is used as the binary response in section 5.

2.2. Functional predictor: intraday returns

There are 4 trading hours for continuous auction in one typical trading day in Chinese stock market, from 9:30am to 11:30am and 13:00pm to 15:00pm. To depict the intraday returns of SSEC, we use the last price of every five minutes in order to calculate the intraday

7

percentage changes. Denote the price of SSEC at time t on day i as $p_{i,t}$, then the intraday return of SSEC at time t on day i is $r_{i,t} = (p_{i,t} - p_{i,t-1})/p_{i,t-1}$, as commonly defined in most financial studies. Starting from 9:35am to 11:30am and 13:00pm to 15:00pm, 49 points are included in one observation (one trading day). The series of intraday returns are treated as functional data because they could provide the consecutive information on trending of the index as a curve. The relative techniques to deal with functional data are introduced in section 3.1.2.

As there are 345 trading days in our sample, 345 curves of the intraday returns are included in the study, one for each trading day, denoted as $x_i(t), i = 1, ..., n$. $x_i(t)$ is the functional predictor in section 5.

2.3. Scalar predictor: volume

Volume is an important indicator in stock analysis as it is used to measure the relative worth of a market move (Foster and Viswanathan, 1993; Lillo et al., 2003). Denote the volume on day i in Shanghai Stock Exchange as $z_i, i = 1, ..., n$ and we use it as the scalar predictor in section 5.

The data introduced in section 2.1, 2.2, 2.3, are downloaded from Thomson Reuters' Tick History.

2.4. Compositional predictor: market emotion

Recent studies have found that the relationship between social media sentiments and stock returns is time-varying (Ho et al., 2017) and some have successfully incorporated investors' emotions from social media into predicting stocks' prices (Zhou et al., 2017b; Sun et al., 2017; Ruan et al., 2018; Li et al., 2014). In this paper, we use the emotion measures from Zhou et al. (2017b). Based on over 3.5 million emotionally labelled tweets from Sina Weibo as the corpus and a fast Naive Bayes classifier (Zhao et al., 2012), they have arranged the daily stock-relevant tweets into five categories, namely "anger", "disgust",

8

"joy", "sadness", and "fear". By scaling each kind of emotion by the sum of tweets in a day, we obtain the daily emotion ratios, respectively, to represent the investors' emotion ratios towards the market. The data are naturally compositional, each observation (one trading day) has five parts and the five parts add up to 1, denoted as $c_i = [c_{i1}, c_{i2}, c_{i3}, c_{i4}, c_{i5}], i = 1, ..., n$. Fig. 2 illustrates the daily market emotion ratios over the sample period. The "fear" and "joy" had kept been the two dominant feelings among the five types of market emotion. Specifically, in the first phase, "joy" is the largest part of the market emotions and "fear" is the second. In the second phase, however, the two kinds of emotions switch their positions, "fear" became to be the largest. In the third phase, "joy" and "fear" alternatively dominates the market emotions. In addition, "fear" began to grow while "disgust" shrank a little in the bearish market (phase 2 and phase 3). The dominant roles of "fear" and "joy" in the compositions, however, does not mean that they are consequentially of the most indicative to the market trend. Section 5 will discuss this issue in details.
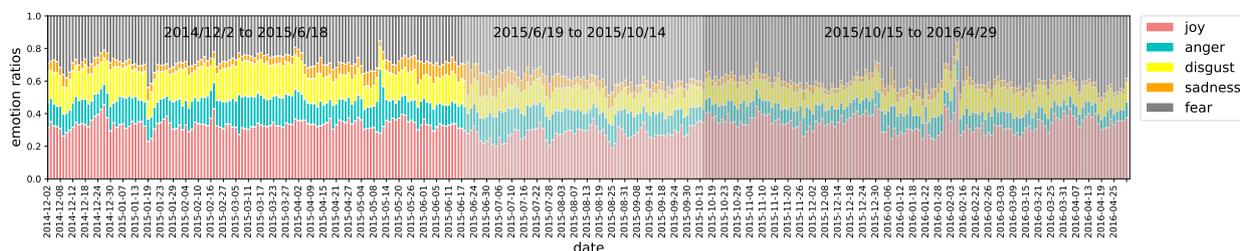

Figure 2: **Market emotions from Sina Weibo (Zhou et al., 2017b).** The sample period is divided into three phases illustrated by different backgrounds, as in Fig. 1.

3. Methodology

In this section, we first introduce how to deal with compositional data and functional data using transformation techniques. Then, a model-independent framework for forecasting stock market is proposed based on these preliminaries. Finally, we use logistic regression model as a case of the framework and present the corresponding estimation procedure.

3.1. Preliminaries

3.1.1. Compositional data

One observation of the compositional data is usually represented in the D-part simplex

$$\mathcal{S}^D = \{\boldsymbol{c_i} = [c_{i1}, ..., c_{iD}] \in \mathbb{R}^D | \sum_{j=1}^{D} c_{ij} = 1, 0 < c_{ij} < 1, \forall j\}. \tag{2}$$

Since the natural constraints of compositional data, employing standard linear regression analysis for compositional data often leads to undesirable properties (Aitchison, 1986; Pawlowsky-Glahn et al., 2015). A general solution is to remove the constraints first, and then apply classical statistics analysis on the transformed data to obtain the estimated coefficients. The last step is to transform the estimated coefficients back into the original simplex space. Under this circumstance, the key issue concerns about how to remove the constraints of compositional data by some transformation techniques before building the model. Many efforts have been devoted to investigate such kind of approaches. For instance, the additive logratio transformation (Aitchison, 1986), the centered logratio transformation (Aitchison, 1986), and the isometric logratio (ilr) transformation (Egozcue et al., 2003). Given the fact that ilr transformation is an isometry between $\mathcal{S}^D$ and $\mathbb{R}^{D-1}$, we use it to deal with compositional data in this paper.

For any $\boldsymbol{c_i} = [c_{i1}, c_{i2}, \cdots, c_{iD}] \in \mathcal{S}^D$, the ilr transformation maps $\boldsymbol{c_i}$ to $\boldsymbol{c_i^*} = \mathrm{ilr}(\boldsymbol{c_i}) = (c_{i1}^*, c_{i2}^*, \cdots, c_{i,D-1}^*) \in \mathbb{R}^{D-1}$ by

$$\boldsymbol{c_i^*} = \left(\ln \frac{c_{i1}}{\sqrt[D]{\prod_{j=1}^{D} c_{ij}}}, \ln \frac{c_{i2}}{\sqrt[D]{\prod_{j=1}^{D} c_{ij}}}, \cdots, \ln \frac{c_{iD}}{\sqrt[D]{\prod_{j=1}^{D} c_{ij}}} \right) \boldsymbol{\Psi}', \tag{3}$$

where $\boldsymbol{\Psi}$ is a $(D-1, D)$ matrix

$$
\Psi_{ij} = \begin{cases} +\sqrt{\frac{1}{(D-i)(D-i+1)}}, & j \le D - i, \\ -\sqrt{\frac{D-i}{D-i+1}}, & j = D - i + 1, \\ 0, & otherwise, \end{cases} \tag{4}
$$

as Egozcue et al. (2003) proposed.

3.1.2. Functional data

In recent years, FDA has been rapidly developed (Ramsay and Silverman, 1997, 2007; Ferraty and Vieu, 2006; Horváth and Kokoszka, 2012). When it comes to functional linear regression, for the case of a continuous scalar response variable and a functional predictor for the individual of interest (Ramsay and Silverman, 1997), both non-data-driven basis (e.g. B-spline) and data-driven basis (e.g. functional principal component) are commonly used. Particularly, Hall et al. (2007) considers the least square estimator for functional linear regression model based on functional principal components and obtains the optimal convergence rate of the slope function. Meng et al. (2016) pointed out that functional principal component basis could be the first preference, especially as we have none prior knowledge on the functional data types. Thus, we consider dealing with functional variables based on functional principal components basis expansion.

Define the covariance function of $X(t)$ by $K(s,t) = \text{cov}[X(s), X(t)] = E[X(s)X(t)]$, then by Mercer's Theorem we can obtain the spectral decomposition

$$
K(s,t) = \sum_{j=1}^{\infty} \theta_j \phi_j(s) \phi_j(t), \tag{5}
$$

where $\theta_1 \ge \theta_2 \ge \cdots \ge 0$ are eigenvalues of the operator associated with $K(s,t)$, and $\{\phi_j\}$ are the corresponding eigenfunctions. According to the Karhunen-Loeve representation, we

have $X(t)$ in the space of $\{\phi_j\}_{j=1}^{\infty}$ as

$$X(t) = \sum_{j=1}^{\infty} a_j \phi_j(t). \tag{6}$$

Assume that we have independently identically distributed (i.i.d.) observations $\boldsymbol{X}(t) = (x_1(t), \cdots, x_n(t))$, where n denotes the sample size. Recall that $\boldsymbol{X}(t)$ is assumed to be zero-mean, the empirical covariance function is

$$\hat{K}(s,t) = \frac{1}{n} \sum_{i=1}^{n} x_i(s) x_i(t), \tag{7}$$

which can be used to estimate $K(s,t)$. Same with Equation (5) we can obtain that

$$\hat{K}(s,t) = \sum_{j=1}^{\infty} \hat{\theta}_j \hat{\phi}_j(s) \hat{\phi}_j(t), \tag{8}$$

where $\hat{\theta}_1 \geq \hat{\theta}_2 \geq \cdots \geq 0$. Usually $\int \hat{\phi}_j(t) \hat{\phi}_j(t) dt > 0$ is assumed to get rid of uncertainty of signs. Considering $\{\hat{\phi}\}$ is a basis of the spanned space by $\boldsymbol{X}(t)$, we see that at most n eigenvalues are strictly positive. Then we obtain

$$\boldsymbol{X}(t) \approx \sum_{j=1}^{M} \boldsymbol{a}_j \phi_j(t), \tag{9}$$

where M is the number of the basis functions, and M is usually determined by

$$M = \underset{1 \leq m \leq n}{\arg\min} \{ \sum_{i=1}^{m} \hat{\theta}_i / (\sum_{i=1}^{n} \hat{\theta}_i) \geq \lambda \}, \tag{10}$$

where λ is usually set to be 85% (Wang et al., 2016). From Equation (9), the functional subspace $\mathcal{L}^2$ is spanned by the set of orthonormal basis $\{\hat{\phi}\}$ to M-dimensional real space and accordingly $(\boldsymbol{a}_1, \boldsymbol{a}_2, ..., \boldsymbol{a}_M)$ can be used to represent $\boldsymbol{X}(t)$.

3.2. The framework of aggregating multiple types of complex data

Based on the above, we propose a framework for stock prediction using mixed types of complex data, as shown by Fig. 3. The massive emotional information embedded in social media could be converted into proportions of different kinds of feelings and integrated into the framework as compositional data. The intraday returns could be viewed as curves and delivered to the framework as functional data. Other attributes like daily trading volumes are scalar data. The isometric logratio transformation, functional principal component basis expansion are used to reconstruct the original complex data to provide equivalent numerical transformed data for further statistical analysis. Then, the logistic regression classifier is trained with the transformed data, and the prediction model is thus built. Note that based on the transformed data, any model, either the regression ones or machine learning ones, can be trained to perform the prediction. That's to say, our framework is model-independent and offers an interface of aggregating multiple types of complex data. The prediction model could be used to predict the newly arrived data and give its opinion of whether the market will go up or down.

It is worth noting that the transformation procedures is essentially space transformation, that is, from simplex to real space for compositional data, from Hilbert space to real space for functional data. The isometric logratio transformation fully retains the information of the original compositional data, while the basis function expansion based on functional principal components basis expansion absorbs the most informative elements from the functional data. Therefore, the transformed data well represent the original data on the whole.

Furthermore, suppose that a data set has N samples, each of the functional data sample has p observation points, and each of the compositional data sample has D parts. Then the time complexity of basis function expansion based on functional principal analysis is $\mathcal{O}(Np^2)$ as it is necessary to compute the covariance function in Equation (7) (Meng et al., 2016), while the time complexity of *ilr* transformation is $\mathcal{O}(ND^2)$ as it transforms the
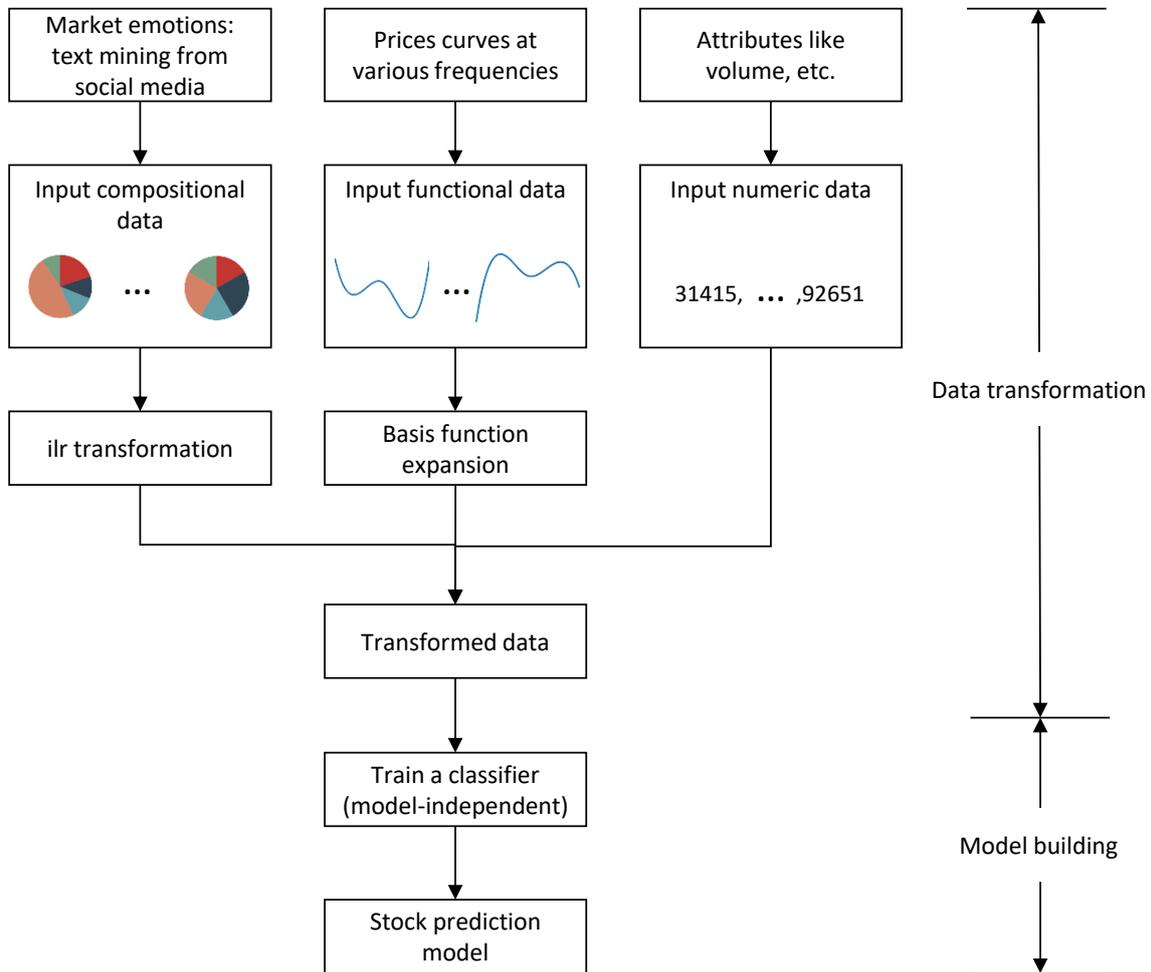
Figure 3: **The framework for stock prediction using mixed types of complex data.**

compositional data sample one-by-one using a matrix of $D \times (D - 1)$ dimensions as shown in Equation (4). Considering that usually $p < 100$ and $D < 10$ in real-world applications, in the big data era, the cost for these transformations is rather small. From this perspective, the framework offers a simple but effective solution to deal with aggregated, big, complex data in the finance domain.

More importantly, the framework is model-independent. It offers an interface of aggregating multiple types of complex data. Based on the transformed data, any kind of

prediction models can be accordingly trained. Classification methods except logistic regression are good alternatives in the model building stage, but we choose logistic regression for the rest of discussion. The reason is that logistic regression not only is particularly useful when the class is dichotomous but also provides the predictors' coefficients, which enables us to collect the intuitions buried in the model.

3.3. Logistic regression and estimation for multiple types of complex data

In this section, we consider a logistic regression model with three kinds of predictors, including scalar data, functional data and compositional data. It is worth noticing that for classifiers like support vector machine, the compositional and functional data are feed into the model in their transformed forms by Equation (3) and (9). However, as logistic regression model involves the coefficients estimation, the inner products of coefficients and variables are required. Therefore, we present the details in the rest of the section.

For the sake of convenience, assume all the variables are zero-mean. Using the inner product expressions for compositional variable and functional variable, our model is

$$Y = h(\gamma Z + \langle \boldsymbol{\alpha}, \boldsymbol{C} \rangle_{\mathcal{S}} + \int X(t)\beta(t)dt + \epsilon), \tag{11}$$

where $Z \in \mathbb{R}$ is a scalar variable, $\boldsymbol{C} \in \mathcal{S}^D$ is a compositional variable of D parts, $X(t) \in \mathcal{L}^2$ is a zero mean, second-order stochastic process, $Y \in \{0, 1\}$ is the binary response. $\langle \cdot, \cdot \rangle_{\mathcal{S}}$ is the corresponding inner product operator for the compositional variable. $\gamma \in \mathbb{R}, \boldsymbol{\alpha} \in \mathcal{S}^D$ and $\beta(t) \in \mathcal{L}^2$ are coefficients to estimate and ϵ is a random error term. As logistic regression takes into account features' joint effects and levels a good linear combination of features as the decision boundary, we consider $h(u) = \frac{\exp(u)}{1+\exp(u)}$ as the link function in this paper.

Note that, the dimension of γ is fixed once Z is given, while $\boldsymbol{\alpha}$ and $\beta(t)$ may be varying under different transformation methods. In the following subsections, we will deal with the compositional and functional data one by one, trying to transform them into scalar data so

as to conduct regular statistics techniques. At the end of the section, we give the maximum likelihood estimation procedure using the transformed data. Hereafter, denote sample i as $(y_i, z_i, \boldsymbol{c_i}, x_i(t)), i = 1, 2, ..., n$.

According to section 3.1.1, by the ilr transformation, the inner product of $\boldsymbol{\alpha}, \boldsymbol{c_i} \in \mathcal{S}^D$ is converted into

$$\langle \boldsymbol{\alpha}, \boldsymbol{c_i} \rangle_{\mathcal{S}} = \sum_{j=1}^{D-1} \alpha_j^* c_{ij}^*, \tag{12}$$

where $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, ..., \alpha_D] \in \mathcal{S}^D$, $\boldsymbol{\alpha^*} = \mathrm{ilr}(\boldsymbol{\alpha}) = (\alpha_1^*, \alpha_2^*, ..., \alpha_{D-1}^*) \in \mathbb{R}^{D-1}$. Denote $\mathcal{C}(\cdot)$ as the closure for a composition to rescaling of the initial vector so that the sum of its components is 1, i.e. $\mathcal{C}(\boldsymbol{u}) = [\frac{u_1}{\sum_{i=1}^D u_i}, \frac{u_2}{\sum_{i=1}^D u_i}, ..., \frac{u_D}{\sum_{i=1}^D u_i}]$. $\boldsymbol{\alpha^*} \in \mathbb{R}^{D-1}$ could be transformed back to $\boldsymbol{\alpha} \in \mathcal{S}^D$ by the ilr^{-1} transformation

$$\boldsymbol{\alpha} = \mathrm{ilr}^{-1}(\boldsymbol{\alpha^*}) = \mathcal{C}(\exp(\boldsymbol{\alpha^*}\boldsymbol{\Psi})). \tag{13}$$

According to section 3.1.2, we have $X(t)$ and $\beta(t)$ in the space of $\{\phi_j\}_{j=1}^{\infty}$, i.e. $\beta(t) = \sum_{j=1}^{\infty} b_j \phi_j(t)$, similar to Equation (6). Then, considering the orthogonal of the $\phi_j(t)$, the third term in Model (11) can be rewritten as

$$\int X(t)\beta(t)dt = \sum_{j=1}^{\infty} a_j b_j. \tag{14}$$

Again, we assume that $\boldsymbol{X}(t) = (x_1(t), \cdots, x_n(t))$ are independently identically distributed (i.i.d.) observations with zero-mean. By incorporating Equation (8), the group of expansion basis are obtained, and similar to Equation (9), we have

$$\beta(t) = \sum_{j=1}^{M} b_j \phi_j(t), \tag{15}$$

where M is the number of the basis functions, and is determined by Equation (10).

Thus,

$$\int \boldsymbol{X}(t)\beta(t)dt \approx \sum_{j=1}^{M} \boldsymbol{a}_j b_j. \tag{16}$$

Thus, Model (11) can be rewritten as

$$Y \approx h(\gamma Z + \sum_{j=1}^{D-1} \boldsymbol{\alpha}_j^* \boldsymbol{c}_j^* + \sum_{j=1}^{M} \boldsymbol{a}_j b_j + \epsilon), \tag{17}$$

in which parameters to be estimated contain $\gamma, \boldsymbol{\alpha}^*, \boldsymbol{b}$.

For any sample i, let $\boldsymbol{b} = (b_1, \cdots, b_M)'$, $\boldsymbol{a_i} = (a_{i1}, \cdots, a_{iM})'$, $y_i \approx h(\gamma z_i + \boldsymbol{c}_i^{*'}\boldsymbol{\alpha}^* + \boldsymbol{a}_i'\boldsymbol{b} + \epsilon_i)$. The link function is set to be $h(u) = \frac{\exp(u)}{1+\exp(u)}$. For any $u \in \mathbb{R}$. The expectation of Y is

$$\pi_i = P(Y = 1 | z_i, \boldsymbol{c_i}, x_i(t)) = \frac{\exp(\gamma z_i + \boldsymbol{c}_i^{*'}\boldsymbol{\alpha}^* + \boldsymbol{a}_i'\boldsymbol{b} + \epsilon_i)}{1 + \exp(\gamma z_i + \boldsymbol{c}_i^{*'}\boldsymbol{\alpha}^* + \boldsymbol{a}_i'\boldsymbol{b} + \epsilon_i)}, i = 1, ..., n. \tag{18}$$

The likelihood of is $L = \prod_{i=1}^{n} \pi_i^{y_i}(1 - \pi_i)^{1-y_i}$ and we can obtain the estimators by maximum the log likelihood

$$(\hat{\gamma}, \hat{\boldsymbol{\alpha}}^*, \hat{\boldsymbol{b}}) = \arg\max_{\gamma, \hat{\boldsymbol{\alpha}}^*, \hat{\boldsymbol{b}}} \{ \sum_{i=1}^{n} y_i(\gamma z_i + \boldsymbol{c}_i^{*'}\boldsymbol{\alpha}^* + \boldsymbol{a}_i'\boldsymbol{b}) - \sum_{i=1}^{n} ln(1 + \exp(\gamma z_i + \boldsymbol{c}_i^{*'}\boldsymbol{\alpha}^* + \boldsymbol{a}_i'\boldsymbol{b})) \}. \tag{19}$$

The estimated $\hat{\boldsymbol{\alpha}}^*, \hat{\boldsymbol{b}}$ would then be inversely transformed back to their original space denoted as $\hat{\boldsymbol{\alpha}}$ and $\hat{\beta}(t)$ using Equations (13) and (15).

4. Effectiveness of the framework

In this section, we perform simulation studies to verify the effectiveness of the proposed framework with finite sample size. Although the framework serves as an interface between multiple types of complex data and stock prediction, it is important to assess the effectiveness of the parameters estimated by the framework because the parameters could equip the framework with explanatory power.

Given that the data transformation is independent to the prediction model in the framework, the logistic regression is considered as a case of the classification methods to evaluate the robustness and usefulness of the framework, as discussed in section 3.3. The details of the simulation are described as follows.

There are three types of predictors on each individual of interest, i.e., scalar data predictor, compositional data predictor, and functional data predictor. For simplicity, all the predictors are scaled by their centers in order to be zero-mean. The data are generated from the following model

$$y_i = h(\gamma z_i + \langle \boldsymbol{\alpha}, \boldsymbol{c}_i \rangle_{\mathcal{S}} + \int \beta(t) x_i(t) dt + \sigma e_i), \tag{20}$$

where $h(u) = \frac{\exp(u)}{1+\exp(u)}$ is the link function, y_i is the 0-1 response, z_i is the scalar predictor, $\boldsymbol{c}_i$ is the compositional predictor, $x_i(t)$ is the functional predictor, e_i is the noise.

In the simulation, we first generate the predictors. $z_i \in \mathbb{R}$ is generated from normal distribution with mean equal to 0 and standard deviation equal to 1. $\boldsymbol{c}_i$ is of compositional data with three parts and each part is uniformly distributed. e_i is normally distributed. σ controls the ratio of signal to noise and here we set $\sigma = 0.2$, 0.4, and 0.6. Besides, the functional data $x_i(t)$ and its functional coefficients $\beta(t)$ are generated on the $T = 100$ equally spaced grids on $[0, 1]$ as (Hall et al., 2007)

$$\beta(t) = \sum_{j=1}^{50} \beta_j \phi_j(t),$$

$$\beta_1 = 0.3, \beta_j = 4(-1)^{j+1} j^{-2}, j \geq 2,$$

$$\phi_j(t) = \sqrt{2} \cos(j\pi t),$$

$$x_i(t) = \sum_{j=1}^{50} \gamma_j Z_j \phi_j(t),$$

$$\gamma_j = (-1)^{j+1} j^{-a/2}, a = 1.1, Z_j \sim U[-\sqrt{3}, \sqrt{3}].$$

Without loss of generality, let $\gamma = 1, \boldsymbol{\alpha} = [\alpha_1, \alpha_2, \alpha_3] = [0.3, 0.5, 0.2] \in \mathcal{S}^3$. Then the probabilities are calculated by

$$\pi_i = \frac{\exp(\gamma z_i + \langle \boldsymbol{\alpha}, \boldsymbol{c}_i \rangle_\mathcal{S} + \int \beta(t) x_i(t) dt + \sigma e_i)}{1 + \exp(\gamma z_i + \langle \boldsymbol{\alpha}, \boldsymbol{c}_i \rangle_\mathcal{S} + \int \beta(t) x_i(t) dt + \sigma e_i)}, i = 1, ..., n. \tag{21}$$

And we finally obtain n values of the response y_i by simulating observations of a Bernouilli distribution with probabilities π_i.

After generating the simulated data, we use the proposed estimated procedure to obtain the estimated value of $\hat{\gamma}, \hat{\boldsymbol{\alpha}}, \hat{\beta}(t)$. Here we set $\lambda = 85\%$ in Equation (10) to determine the number of basis. The "generate data-estimate coefficients" procedure is repeat 200 times for every sample size setting, i.e. $n = 100, 200, 500, 1000, 2000, 5000, 10000$.

To measure the performances of the estimation procedure, we introduce mean of integrated square error (MISE) and correlation of true $\beta(t)$ and $\hat{\beta}(t)$ as

$$MISE(\hat{\beta}) = \int_0^1 (\hat{\beta}(t) - \beta(t))^2 dt,$$

$$cor(\hat{\beta}(t), \beta(t)) = \frac{\sum_{i=1}^T (\beta(t)_i - \overline{\beta}(t))(\hat{\beta}(t)_i - \overline{\hat{\beta}}(t))}{\sqrt{(\sum_{i=1}^T (\beta(t)_i - \overline{\beta}(t))^2}\sqrt{(\sum_{i=1}^T (\hat{\beta}(t)_i - \overline{\hat{\beta}}(t)))^2}}$$

where $T = 100$ is the number of equally spaced grids on $[0, 1]$. The results of the averaged $MISE(\hat{\beta}(t))$ and $cor(\hat{\beta}(t), \beta(t))$ over the 200 times of simulation are shown in Table 1 and Table 2.

From Table 1, we can see that the estimated parameters are indeed highly correlated with the true parameter. The MISE also shows the unbiased attribute of the estimation procedure in an aggregated level. Both the standard deviations of the two decrease when the sample size becomes larger. As for $\hat{\gamma}, \hat{\boldsymbol{\alpha}}$, their averaged bias and standard deviations are shown in

Table 1: The correlation of estimated functional coefficients $\hat{\beta}(t)$ and true functional coefficients $\beta(t)$ and its standard deviation (in parentheses).

sample size	100	200	500	1000	2000	5000	10000
$\sigma = 0.2$	0.943 (0.049)	0.958 (0.03)	0.962 (0.019)	0.964 (0.012)	0.967 (0.009)	0.967 (0.006)	0.967 (0.004)
$\sigma = 0.4$	0.94 (0.05)	0.952 (0.03)	0.964 (0.017)	0.965 (0.012)	0.966 (0.008)	0.967 (0.005)	0.967 (0.004)
$\sigma = 0.6$	0.937 (0.048)	0.951 (0.032)	0.962 (0.018)	0.966 (0.013)	0.965 (0.009)	0.966 (0.005)	0.967 (0.004)

Table 2: $MISE(\hat{\beta})$: the MISE of the functional coefficients and its standard deviations (in parentheses).

sample size	100	200	500	1000	2000	5000	10000
$\sigma = 0.2$	0.263 (0.206)	0.16 (0.109)	0.122 (0.058)	0.109 (0.035)	0.099 (0.026)	0.098 (0.016)	0.097 (0.012)
$\sigma = 0.4$	0.263 (0.193)	0.172 (0.097)	0.117 (0.052)	0.108 (0.036)	0.104 (0.026)	0.1 (0.017)	0.101 (0.011)
$\sigma = 0.6$	0.287 (0.486)	0.173 (0.111)	0.126 (0.053)	0.112 (0.039)	0.112 (0.028)	0.108 (0.017)	0.106 (0.012)

Table 3 and Table 4, which both exhibit only tiny bias. Looking at the estimation variance of the bias, we can see it does decrease when we increase the sample size. In conclusion, the simulation study above has provided unbiased and consistent estimation results, which advertises the effectiveness of the proposed framework.

5. Empirical studies on Chinese stock market

Considering the low cost and evaluated consistency of the framework in aggregating complex data, it can be used in realistic applications like stock prediction. As the Chinese stock market's fundamentals vary throughout the sample period, we cut the whole period into three phases, as shown in Fig. 1. We use the volume, emotion, and intraday returns of day i as the predictors, while the states of the open return (if the return is positive, then $y_i = 1$, otherwise $y_i = 0$) in day $i+1$ as the response. In this case, the framework serves as a prediction method for the open return of the index. Here, when transforming the functional data, λ is set to be 99% in Equation (10) to determine the number of basis so as to retain the majority of information that the intraday return series could offer. Furthermore, note that unlike machine learning methods like support vector machine, the advantage of logistic

Table 3: The bias of scalar predictor's coefficient $\hat{\gamma}$ and its standard deviations (in parentheses).

sample size	100	200	500	1000	2000	5000	10000
$\sigma = 0.2$	0.083 (0.39)	0.041 (0.271)	−0.005 (0.154)	−0.024 (0.113)	−0.036 (0.069)	−0.041 (0.048)	−0.035 (0.029)
$\sigma = 0.4$	0.074 (0.391)	−0.016 (0.263)	−0.036 (0.149)	−0.031 (0.103)	−0.043 (0.075)	−0.047 (0.046)	−0.057 (0.035)
$\sigma = 0.6$	0.004 (0.413)	−0.005 (0.265)	−0.058 (0.153)	−0.072 (0.102)	−0.076 (0.066)	−0.085 (0.048)	−0.085 (0.031)

Table 4: The bias of compositional coefficients $\hat{\boldsymbol{\alpha}}$ and its standard deviations (in parentheses).

sample size		100	200	500	1000	2000	5000	10000
$\sigma = 0.2$	$\hat{\alpha}_1$	0.001 (0.1)	0 (0.056)	−0.001 (0.034)	0.003 (0.024)	0.001 (0.017)	0.002 (0.011)	0.001 (0.008)
	$\hat{\alpha}_2$	0 (0.105)	−0.002 (0.073)	0.002 (0.043)	−0.006 (0.028)	−0.003 (0.021)	−0.005 (0.013)	−0.005 (0.009)
	$\hat{\alpha}_3$	−0.001 (0.072)	0.002 (0.044)	−0.001 (0.028)	0.003 (0.019)	0.002 (0.014)	0.003 (0.008)	0.004 (0.006)
$\sigma = 0.4$	$\hat{\alpha}_1$	−0.003 (0.091)	−0.006 (0.06)	0.003 (0.034)	0.004 (0.022)	0.001 (0.017)	0.002 (0.011)	0.002 (0.007)
	$\hat{\alpha}_2$	0.011 (0.113)	0.007 (0.074)	−0.004 (0.044)	−0.01 (0.03)	−0.006 (0.022)	−0.008 (0.012)	−0.009 (0.009)
	$\hat{\alpha}_3$	−0.008 (0.07)	−0.001 (0.047)	0.001 (0.031)	0.006 (0.021)	0.005 (0.015)	0.006 (0.009)	0.006 (0.006)
$\sigma = 0.6$	$\hat{\alpha}_1$	0.013 (0.1)	0.005 (0.055)	0.006 (0.035)	0.003 (0.026)	0.003 (0.016)	0.004 (0.011)	0.004 (0.007)
	$\hat{\alpha}_2$	−0.014 (0.111)	−0.012 (0.069)	−0.017 (0.039)	−0.013 (0.031)	−0.012 (0.021)	−0.014 (0.013)	−0.013 (0.009)
	$\hat{\alpha}_3$	0.001 (0.072)	0.007 (0.051)	0.011 (0.028)	0.01 (0.019)	0.009 (0.013)	0.01 (0.009)	0.009 (0.006)

regression is that it could access the coefficients of the variables instead of focusing on the prediction power alone. As both the interpretation and the predicting power of the proposed framework are of interest, we present them one by one in the following sub-sections.

5.1. Predication

We perform the 5-fold cross validation for the three phases separately. The cross validation is a widely used method to estimate how accurately a predictive model will perform in practice. In a 5-fold cross-validation, the original sample is randomly split into 5 equal size sub-samples. Of the 5 sub-samples, a single sub-sample is left as the validation data for testing the model, and the remaining 4 sub-samples are used as training data. The cross-validation process is then repeated 5 times, with each of the 5 sub-samples used exactly once

for the validation. The 5 results from the folds is averaged to produce a single estimation of accuracy. The advantage of this method is that all observations are used for both training and validation, and each of them is used for validation only once.

In the benchmark scenario, we convert the daily open return into binary response by using zero as the cut-off point, as discussed in the data section. That is, if the open return is positive, $y_i = 1$, otherwise $y_i = 0$. The accuracy is defined as the rate of observations correctly classified using 0.5 as the cut point of the predicted probability π.

The prediction accuracies are shown in Table 5. The last two rows of Table 5 show the result of the comparison experiments, in which the original data (functional data, compositional data, scalar data) are regarded as scalar predictors and are fed into the classification model. As can be seen, the prediction accuracies using the original data are indeed worse than those under our proposed framework. As a matter of fact, treating the functional data and compositional data as scalar predictors and directly imposing them together on statistic analysis can not be supported by statistical theory because they belong to different spaces.

Using the proposed framework, we have the accuracy of 0.65, 0.65, and 0.56 for the three phases, respectively. A substitution of the logistic regression is support vector machine (SVM), and the second row of Table 5 shows the prediction accuracies that SVM provides. As can be seen, the SVM classifier (using a linear kernel) does not out-perform the logistic regression under the proposed framework. This result coincides with the previous study (Perlich et al., 2003), in which the analysis of learning curves shows that logistic regression performs well for small data sets. Due to advantages of the good prediction accuracies and the convenience for interpretation of regression coefficients, the following discussion focus on the results that logistic regression yields.

The accuracies for the first two phases are higher than the previous study (Zhou et al., 2017b), in which the logistic regression is also used based on only the emotion data and obtain an accuracy of 58.1%, while the third phase failed to out-perform the previous study.

Table 5: Prediction accuracy.

	classification model	phase 1	phase 2	phase 3
under the proposed framework	logistic regression	**0.65**	**0.65**	**0.56**
	SVM	0.65	0.59	0.50
using original data	logistic regression	0.54	0.44	0.46
	SVM	0.50	0.57	0.47

The reason might be that in phase 3, the market kept being depressed and less information could be extracted from the variables we have. Nonetheless, the results have shown a good prediction capacity of the proposed framework based on multiple types of complex data.

At first glance of Table 5, the prediction accuracies of the proposed framework may not seem to be exciting. This is most likely due to the fact that, in the present modelling, we are using zero as the cut-off point of the daily open return, which is too sensitive to capture the significant difference between the ups and downs of the market. To weaken the sensitivity, we further present a threshold-based sampling approach. Define τ as the cut-off point of the daily open return, where $\tau \in [0, 0.1]$. For any given τ, we pick out the observation where the daily open return is higher than τ or lower than $-\tau$ and then perform the 5-fold cross validation of the prediction framework. Fig. 4 illustrates the relationship between τ and the accuracy. The maximum accuracies are also shown in the figures, with the corresponding sample sizes n and thresholds τ. As can be seen, the accuracy is sensitive to the threshold τ. The trade-off here is that τ has successfully left the returns that of small absolute values out of consideration, while the number of observations is cut down, making the prediction accuracy unstable. Thus, the relationship between τ and accuracy is not monotonic. Nevertheless, Fig. 4 exhibits that the accuracy could be improved by selecting the significant ups or downs of the market in advance and sacrificing the number of observations, especially for phase 1 and phase 2, when the market experiencing big ups and downs. From this perspective, the framework is expected to reach better performance when the market vibrates sharply. Besides, Fig. 4 gives a comprehensive evaluation of the

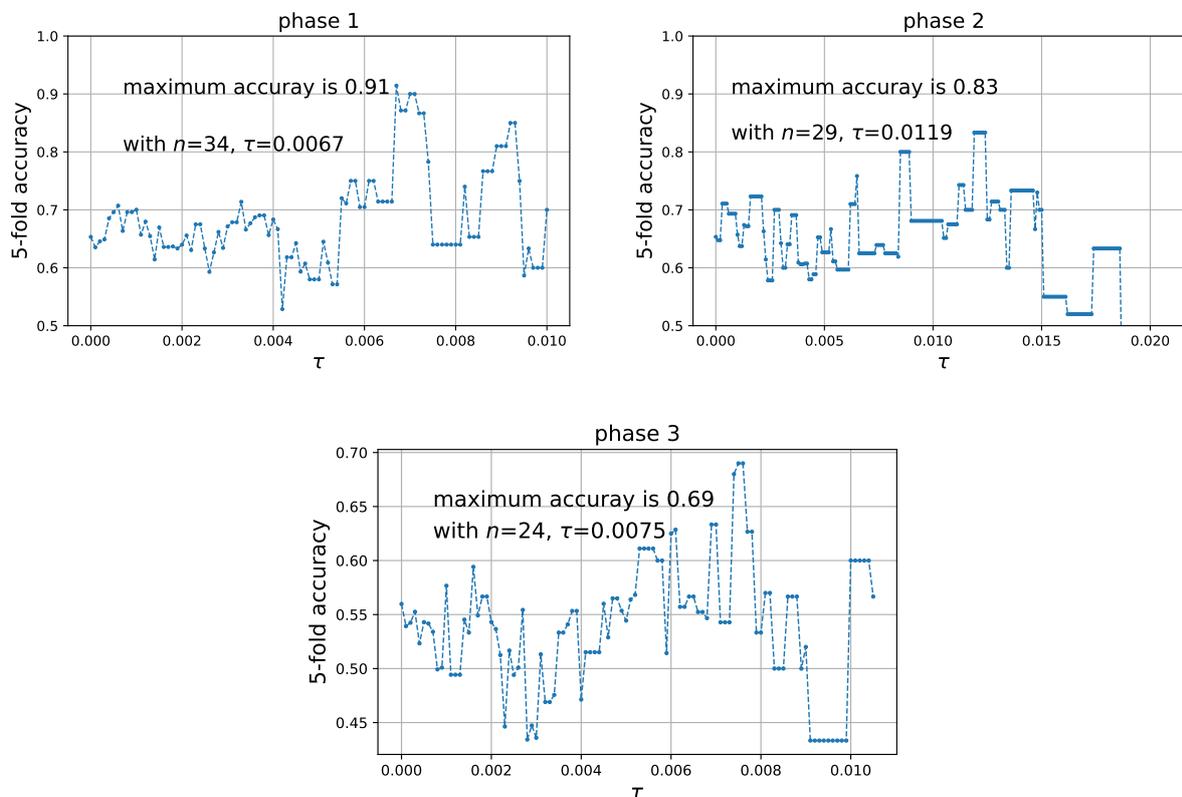predictive power of the proposed framework.

Figure 4: **The relationship between the cut point τ and the 5-fold accuracy.** When $\tau = 0$, the accuracy of is 0.65, 0.65, and 0.56 for the three phases as reported in Table 5.

5.2. Coefficients interpretation

For the three different sample periods, we apply the proposed method and obtain the estimated coefficients of compositional predictors (investors' emotion), functional predictors (series of the intraday 5-minutes returns), and scalar predictor (volume). The estimated coefficients of compositional predictors and scalar predictor are shown in Table 6. As can be seen, in the bullish market (phase 1), the most important emotion is "sadness". Though in a bullish market, the dominant emotion would be "joy" (as shown by Fig. 2), the "sadness" is the one that has the greatest influence on the opening return. In phase 2, the "disgust" turns out to be the most influential one among the five kinds of emotion. This indicates that

24

during the beginning of a bearish market, people's dislike for the market has the absolute priority to the market trending. In phase 3, where the market kept vibrating in depression, however, the "anger" is of great importance to the opening return in the future, while other types of emotion plays their parts at the same time. It is worth noticing that "fear" becomes important in phase 3 instead of phase 2, implying that it is not the fear at the beginning of the bearish market but the fear after the initial shock would strike the market trend.

Table 6: Coefficients of emotions and volume.

	phase 1	phase 2	phase 3
anger	0.01	0.00	**0.49**
disgust	0.04	**0.98**	0.10
joy	0.29	0.00	0.12
sadness	**0.66**	0.00	0.13
fear	0.00	0.02	0.16
volume	-0.84	-0.79	-0.24

Table 6 also shows that the coefficient of volume is always negative. In fact, a high level of volume means the market participants hold diverge opinions on the market expectation, that is, some think it is time to sell while others believe that it is time to buy. Therefore, the negative coefficients imply that one unit higher of volume (or say, one unit of higher divergence of investors' opinions) happened yesterday will lead to some units lower of the probability for the stock index to perform a positive return at opening. And the absolute impact is decreasing from the bullish market to the bearish market.

The estimated coefficients of functional predictors are shown in Fig. 5. The functional coefficients could be explained as the impact of the return at a specific time of yesterday on today's probability of opening return being positive. It illustrates how the intraday effect (Harris, 1986; Chang et al., 2008; Foster and Viswanathan, 1993) of stock prices influence the trending of the market during the boom-and-bust of Chinese stock market.

Specifically, Fig. 5 illustrates the different characteristic of this impact in the three phases. In the bullish market (phase 1), the impact increases to the top and goes down to zero
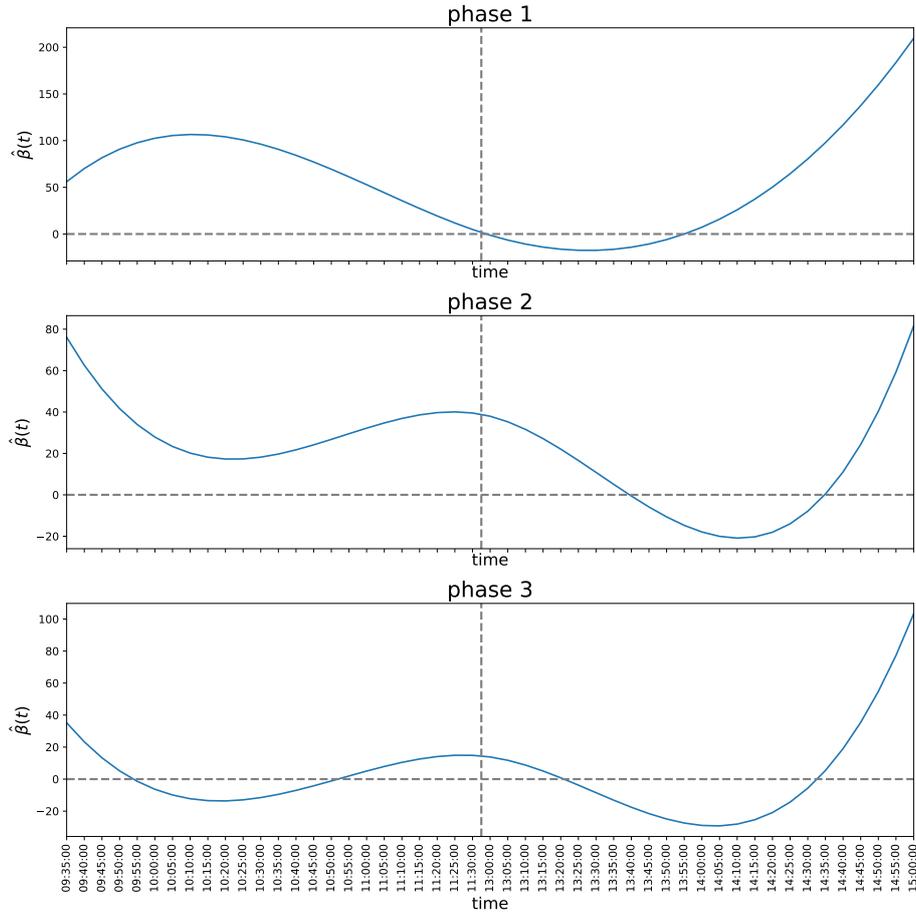
Figure 5: **Functional coefficients in the three phases.** The x-axis is the intraday time at 5 minutes frequency, from 9:35am to 11:30am and 13:00pm to 15:00pm. The curves illustrate the degree of impact of the intraday returns in the former day on the opening return today.

before the closing in the morning. In the afternoon, the impact decreases to below zero at the beginning and goes up before closing.

During the initial market shock (phase 2), the impact of the intraday returns on the next day's probability of a positive open return decreases at the beginning of the day. Different from phase 1, the impact keeps being positive in the morning and the first half hour in the afternoon. It goes down to negative in the afternoon and reverses back again to positive and flies to a high level before closing. The persistent positive impact throughout a day may be induced by the investors's persistent attention towards the market and the intraday returns in turn exert influence to the opening return in the next day.

26

Comparing to phase 1 and phase 2, phase 3 has the least of such kind of impact overall, as the functional coefficient is around zero in most of the time in the morning. This may source from the fact that the depressed state of the market could stir up little influence on the following day. It reaches its lowest around 14:00pm, similar with phase 2. This is just what the investors calls "Magic 14:00" and "Magic 14:30" that has terrified the market participants because the sharp falls always happened around 14:00pm to 14:30pm during the 2015 Chinese market crash. Our results further confirm that those special moments have a high level of negative impact on the following days.

The discussion in this section demonstrates the ability of the proposed framework to integrate multiple types of complex data and to explain how the variables work for the response.

6. Conclusion

In this paper, we proposed a framework that aggregates three types of data forms, namely scalar variable, compositional variable, and functional variable, for predicting stock market. While the framework is model-independent, we mainly inspect the logistic regression in this study. In terms of isometric logratio transformation, functional principal component and logistic regression, we develop the estimation procedure of the framework in aggregating complex data. Numeric simulation experiments show that our proposed framework is effective.

In the empirical studies on Chinese stock market, the trading volume (scalar data), intraday return series (functional data), and investors' emotion (compositional data) from social media are used to predict whether the market is up or down at opening in the next day. By dividing the sample period into 3 phases, we find that the estimated coefficients of trading volume and intraday return series shed light on the different market status. Most surprisingly, we find that in the bullish market, the "sadness" is more important to the future

market trend than "joy". In the initial collapse (phase 2), the "disgust" plays a dominant role. When the market became depressed, the "anger" and "fear" begun to do their parts. Interestingly, our results show that it is not at the beginning of the bearish market but the subsequent period in which the investors' "fear" comes to be indicative to market trend. Besides, our proposed method exhibits a competent prediction power, especially in the first two phases.

Though our proposed framework performs well in our study, it has inevitable limitations. For instance, the correlation among the time series is neglected as we treat the observations independently. Besides, the accuracy is not high enough for phase 3, where there's a depressed market. Future works could consider develop a panel data framework to solve the first limitation, and add other informative variables into the framework to solve the second.

7. Acknowledgments

8. Reference

References

Aebischer, N. J., Robertson, P. A., Kenward, R. E., 1993. Compositional analysis of habitat use from animal radio-tracking data. Ecology 74 (5), 1313–1325.

Aitchison, J., 1986. The statistical analysis of compositional data. Springer, Chapman & Hall, London.

Baralis, E., Cagliero, L., Cerquitelli, T., Garza, P., Pulvirenti, F., 2017. Discovering profitable stocks for intraday trading. Information Sciences 405, 91–106.

Bingham, R. L., Brennan, L. A., Ballard, B. M., 2007. Misclassified resource selection: compositional analysis and unused habitat. Journal of Wildlife Management 71 (4), 1369–1374.

Buccianti, A., Tassi, F., Vaselli, O., 2006. Compositional changes in a fumarolic field, vulcano island, italy: a statistical case study. Geological Society, London, Special Publications 264 (1), 67–77.

Cai, T. T., Hall, P., et al., 2006. Prediction in functional linear regression. The Annals of Statistics 34 (5), 2159–2179.

Chang, S.-C., Chen, S.-S., Chou, R. K., Lin, Y.-H., 2008. Weather and intraday patterns in stock returns and trading activity. Journal of Banking & Finance 32 (9), 1754–1766.

Chayes, F., 1960. On correlation between variables of constant sum. Journal of Geophysical research 65 (12), 4185–4193.

Chen, M.-Y., Chen, B.-T., 2015. A hybrid fuzzy time series model based on granular computing for stock price forecasting. Information Sciences 294, 227–241.

Chen, N.-F., Roll, R., Ross, S. A., 1986. Economic forces and the stock market. Journal of business, 383–403.

Comte, F., Johannes, J., et al., 2012. Adaptive functional linear regression. The Annals of Statistics 40 (6), 2765–2797.

Duan, J., Liu, H., Zeng, J., 2013. Posterior probability model for stock return prediction based on analysts recommendation behavior. Knowledge-Based Systems 50, 151–158.

Efendi, R., Arbaiy, N., Deris, M. M., 2018. A new procedure in stock market forecasting based on fuzzy random auto-regression time series model. Information Sciences 441, 113–132.

Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barcelo-Vidal, C., 2003. Isometric logratio transformations for compositional data analysis. Mathematical Geology 35 (3), 279–300.

Escabias, M., Aguilera, A. M., Valderrama, M. J., 2004. Principal component estimation of functional logistic regression: discussion of two different approaches. Journal of Nonparametric Statistics 16 (3-4), 365–384.

Fan, Y., James, G. M., Radchenko, P., et al., 2015. Functional additive regression. The Annals of Statistics 43 (5), 2296–2325.

Ferraty, F., Vieu, P., 2006. Nonparametric functional data analysis: theory and practice. Springer, New York.

Foster, F. D., Viswanathan, S., 1993. Variations in trading volume, return volatility, and trading costs: Evidence on recent price formation models. The Journal of Finance 48 (1), 187–211.

García-Portugués, E., González-Manteiga, W., Febrero-Bande, M., 2014. A goodness-of-fit test for the functional linear model with scalar response. Journal of Computational and Graphical Statistics 23 (3), 761–778.

Geva, T., Zahavi, J., 2014. Empirical evaluation of an automated intraday stock recommendation system

incorporating both market data and textual news. Decision support systems 57, 212–223.

Godichon-Baggioni, A., Maugis-Rabusseau, C., Rau, A., 2018. Clustering transformed compositional data using k-means, with applications in gene expression and bicycle sharing system data. Journal of Applied Statistics, 1–19.

Hagenau, M., Liebmann, M., Neumann, D., 2013. Automated news reading: Stock price prediction based on financial news using context-capturing features. Decision Support Systems 55 (3), 685–697.

Hall, P., Hooker, G., 2016. Truncated linear models for functional data. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 78 (3), 637–653.

Hall, P., Horowitz, J. L., et al., 2007. Methodology and convergence rates for functional linear regression. The Annals of Statistics 35 (1), 70–91.

Harris, L., 1986. A transaction data study of weekly and intradaily patterns in stock returns. Journal of financial economics 16 (1), 99–117.

Harris, L., 1989. A day-end transaction price anomaly. Journal of Financial and Quantitative Analysis 24 (1), 29–45.

Ho, C.-S., Damien, P., Gu, B., Konana, P., 2017. The time-varying nature of social media sentiments in modeling stock returns. Decision Support Systems 101, 69–81.

Horváth, L., Kokoszka, P., 2012. Inference for functional data with applications. Vol. 200. Springer.

Hsu, C.-M., 2011. A hybrid procedure for stock price prediction by integrating self-organizing map and genetic programming. Expert Systems with Applications 38 (11), 14026–14036.

Huang, L., Zhao, J., Wang, H., Wang, S., 2016. Robust shrinkage estimation and selection for functional multiple linear model through lad loss. Computational Statistics & Data Analysis 103, 384–400.

Jain, P. C., Joh, G.-H., 1988. The dependence between hourly prices and trading volume. Journal of Financial and Quantitative Analysis 23 (3), 269–283.

Jasemi, M., Kimiagari, A. M., Memariani, A., 2011. A modern neural network model to do stock market timing on the basis of the ancient investment technique of japanese candlestick. Expert Systems with Applications 38 (4), 3884–3890.

Jones, C. P., Litzenberger, R. H., 1970. Quarterly earnings reports and intermediate stock price trends. The Journal of Finance 25 (1), 143–148.

Li, Q., Wang, T., Li, P., Liu, L., Gong, Q., Chen, Y., 2014. The effect of news and public mood on stock movements. Information Sciences 278, 826–840.

Lillo, F., Farmer, J. D., Mantegna, R. N., 2003. Econophysics: Master curve for price-impact function.

Nature 421 (6919), 129.

Longford, N. T., Pittau, M. G., 2006. Stability of household income in european countries in the 1990s. Computational statistics & data analysis 51 (2), 1364–1383.

Lu, S., Zhao, J., Wang, H., Ren, R., 2017. Herding boosts too-connected-to-fail risk in stock market of china. arXiv preprint arXiv:1705.08240.

Matías, J. M., Reboredo, J. C., 2012. Forecasting performance of nonlinear models for intraday stock returns. Journal of Forecasting 31 (2), 172–188.

Meng, Y., Liang, J., Qian, Y., 2016. Comparison study of orthonormal representations of functional data in classification. Knowledge-Based Systems 97, 224–236.

Miesch, A., Chapman, R., 1977. Log transformations in geochemistry. Journal of the International Association for Mathematical Geology 9 (2), 191–198.

Pan, Y., Xiao, Z., Wang, X., Yang, D., 2017. A multiple support vector machine approach to stock index forecasting with mixed frequency sampling. Knowledge-Based Systems 122, 90–102.

Pawlowsky-Glahn, V., Egozcue, J. J., Tolosana-Delgado, R., 2015. Modeling and analysis of compositional data. John Wiley & Sons.

Pearson, K., 1896. Mathematical contributions to the theory of evolution. iii. regression, heredity, and panmixia. Philosophical Transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character 187, 253–318.

Perlich, C., Provost, F., Simonoff, J. S., 2003. Tree induction vs. logistic regression: A learning-curve analysis. Journal of Machine Learning Research 4 (Jun), 211–255.

Preis, T., Moat, H. S., Stanley, H. E., 2013. Quantifying trading behavior in financial markets using google trends. Scientific reports 3, srep01684.

Ramsay, J. O., Silverman, B. W., 1997. Functional data analysis. Springer, New York.

Ramsay, J. O., Silverman, B. W., 2007. Applied functional data analysis: methods and case studies. Springer.

Ruan, Y., Durresi, A., Alfantoukh, L., 2018. Using twitter trust network for stock market analysis. Knowledge-Based Systems 145, 207–218.

Shang, Z., Cheng, G., et al., 2015. Nonparametric inference in generalized functional linear models. The Annals of Statistics 43 (4), 1742–1773.

Sun, T., Wang, J., Zhang, P., Cao, Y., Liu, B., Wang, D., 2017. Predicting stock price returns using microblog sentiment for chinese stock market. In: Big Data Computing and Communications (BIGCOM), 2017 3rd International Conference on. IEEE, pp. 87–96.

Wang, H., Huang, L., Wang, S., 2016. Generalized linear regression model based on functional data analysis. Journal of Beijing University of Aeronautics and Astronautics (1), 8–12.

Wanyun, C., Jie, L., 2013. Investors bullish sentiment of social media and stock market indices. J. Manag 5, 012.

Ye, F., Zhang, L., Zhang, D., Fujita, H., Gong, Z., 2016. A novel forecasting method based on multi-order fuzzy time series and technical analysis. Information Sciences 367, 41–57.

Zhang, X., Zhang, Y., Wang, S., Yao, Y., Fang, B., Philip, S. Y., 2017. Improving stock market prediction via heterogeneous information fusion. Knowledge-Based Systems.

Zhao, J., Dong, L., Wu, J., Xu, K., 2012. Moodlens: an emoticon-based sentiment analysis system for chinese tweets. In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp. 1528–1531.

Zhou, L., Lu, D., Fujita, H., 2015. The performance of corporate financial distress prediction models with features selection guided by domain knowledge and data mining approaches. Knowledge-Based Systems 85, 52–61.

Zhou, L., Si, Y.-W., Fujita, H., 2017a. Predicting the listing statuses of chinese-listed companies using decision trees combined with an improved filter feature selection method. Knowledge-Based Systems 128, 93–101.

Zhou, Z., Xu, K., Zhao, J., Sep 2017b. Tales of emotion and stock in china: volatility, causality and prediction. World Wide Web.