Hindawi
Journal of Sensors
Volume 2021, Article ID 1089266, 12 pages
https://doi.org/10.1155/2021/1089266

Research Article

Correlation Analysis of Stocks and PMI Index Based on Logistic Regression Model

Qiong Kang

Henan Industrial Vocational and Technical College, Nanyang 473000, China

Correspondence should be addressed to Qiong Kang; 2009013@hnpi.edu.cn

Received 15 July 2021; Revised 14 August 2021; Accepted 18 August 2021; Published 15 September 2021

Academic Editor: Mu Zhou

In order to explore the correlation between stocks and the PMI index, based on the generalized logistic loss and margin distribution, this paper designs a margin distribution logistic regression model that is easy to optimize, has robustness, and generalization ability, and gives a multiclass margin distribution logistic regression framework. This framework can be used to perform two-classification, multiclassification, and feature selection tasks. Moreover, this paper gives a training algorithm for margin distribution logistic regression on large-scale data sets through the pairwise stochastic gradient descent method. In addition, this paper combines the logistic regression model to construct a correlation analysis model between stocks and PMI index and uses the PMI data of the National Bureau of Statistics as a sample to design experiments to verify the performance of the system model constructed in this paper. From the experimental analysis, it can be seen that the algorithm constructed in this paper has a certain effect, and the strong correlation between PMI and stocks has been further verified.

1. Introduction

At this stage, the international and domestic economic environment is becoming more and more complicated, and there are many uncertain risk factors. Therefore, we need comprehensive and advanced economic indicators to express our country's economic conditions, and the purchasing manager index is just such an indicator [1].

In recent years, the PMI has received widespread attention from government agencies, financial circles, and even ordinary people. After the official PMI is released on the first working day of each month, the major financial media and financial institutions will reprint it as soon as possible and use this as a basis for a reasonable analysis of the economic trend in the future. The release and continuous application of the PMI system in the actual economy are one of the important manifestations of my country's continuous economic development and embracing the world [2]. Therefore, research on the origin of the PMI system is conducive to the continuous progress of my country's PMI system. Official PMI and HSBC PMI are one of the important products of the rapid development of market economy. The influence of the financial market on the international economy is becoming more and more significant, and the stock market, as an extremely important part of the financial market, has developed rapidly in recent years [3]. The volatility of stock prices is one of the universal laws of the stock market, and it is also a key point of concern to the general public and regulatory authorities. Therefore, the accuracy and timeliness of stock price analysis are an important goal for all relevant practitioners. Many authoritative media and security companies pay more attention to the PMI index, and reprint it as soon as the monthly PMI index is released, and use it as an important basis for analyzing the future stock market trend, and the PMI index is also familiar to more ordinary people.

2. Related Work

Through research, the literature [4] found that PMI has a good role in predicting macroeconomic information such as business cycles and economic growth, can provide early clues to the transition of economic development, and can improve the accuracy of prediction; that is, the biggest feature of PMI is its advancement. The literature [5] indicates that economists are very concerned about PMI, especially when the turning point of economic development is approaching. The literature [6] studied the critical value of PMI in prediction. When the purchasing manager index is >47, it generally means that the manufacturing industry is in a state of expansion and the development prospects are optimistic. When the purchasing manager index is >47, the gross national product (GDP) generally maintains a positive growth, and the overall economic situation is good. When PMI > 52.5, it is often accompanied by an increase in short-term interest rates. The literature [7] found that PMI and other related indicators of the manufacturing industry are consistent, which can reflect the overall real situation of the entire manufacturing industry. The literature [8] found through descriptive statistics that the trend of purchasing manager index and GDP growth rate is highly correlated. Moreover, it analyzed the growth rate of PMI and GDP and the growth rate of personal income and found that PMI can improve the prediction accuracy of both in the current period, and the prediction of the next period can improve the prediction accuracy more greatly. The literature [9] used the sum of the production index PMI in the manufacturing and nonmanufacturing purchasing manager index systems and its first-order difference ΔPMI to perform an OLS regression analysis on GDP growth rates and concluded that the single diffusion index in the PMI system can also predict economic growth. The literature [10] carried out a weighted average of the two individual indicators of new orders and supplier delivery time in the PMI system. The research results show that the new indicators are highly correlated with GDP. The literature [11] shows the effectiveness of the Professional Forecast Survey (SRF) and PMI for forecasting actual economic activities.

The literature [12] gave a brief introduction to the establishment background and preparation process of the PMI system and pointed out the important significance of the establishment of the PMI system. The literature [13] proposed to establish a PMI system in line with national conditions to enhance the authority and persuasiveness of the indicators. The literature [14] put forward two suggestions for the establishment and development of the PMI system. The first is to be in line with international standards; to learn the survey methods, sample selection, and calculation methods of the foreign mature and advanced PMI system; and to discuss the advantages of the current PPS sampling method. The second is to continuously improve the system structure based on actual national conditions, ensure the scientificity of sample selection, increase the participation of sample units, and ensure the truth and validity of statistics. The literature [15] made a brief analysis of PMI from three aspects: first, the concept and impact of the PMI system; second, the investigation and calculation method of PMI; and third, the indicative role of PMI on economic operations. The literature [16] studied the relationship between manufacturing purchasing manager indices and obtained relevant conclusions by establishing a VAR model and using impulse response and variance decomposition methods.

The literature [17] carried out a meaningful conversion and reduction of the individual diffusion data in the PMI system and found that PMI is ahead of related indicators such as fixed asset investment (FAI), factory price of industrial products (PPI), and customs import and export and further verifies the predictive effect of PMI on the macroeconomic cycle. Based on the significant impact of total inventory investment on GDP fluctuations, the literature [18] used PMI, an indicator widely used in economic fluctuation forecasting, to analyze the periodicity of finished product inventory investment and raw material inventory investment, and found that the former is counter-cyclical and the latter is vice versa. The literature [19] used the VAR model to conduct Granger causality test and Johnson cointegration test and concluded that PMI is the Granger cause of GDP growth, and there is a long-term equilibrium relationship between the two, and PMI can be used to effectively predict economic growth.

3. Margin Distribution Logistic Regression Model

A classification model is mainly composed of two parts:

$$\text{Loss} = l\left(y, w^T x\right) + r(w). \tag{1}$$

The first part is the loss function l of the classification, and the second part is the regularization term r of the model. When designing a classification model, the robustness of the model and the generalization ability of the model need to be considered. From the perspective of loss function, the huge loss function value caused by unreasonable outliers cannot have an excessive impact on the normal classification loss, which is caused by several unreasonable outliers. With unreasonable classification function, so the robustness of the model should be mainly considered from the perspective of the loss function of the model. From the perspective of the generalization ability of the model, there are often a lot of noise points in real data. These noise samples may be flooded in the vicinity of the two types of classification and discrimination hyperplanes, making it difficult for the model to find the true classification function. And how to resist the influence of these noise points is a key factor in the generalization ability of the model. The regularization technology in the model can help introduce some a priori assumptions, such as soft interval, maximize minimum interval, and optimize interval distribution. In this article, we mainly consider how to build a classification model with robustness and generalization ability from the loss function of the model and the introduction of a priori assumptions, specifically by introducing a smooth, convex, and generalized weakly

sensitive to outliers generalized logistic loss; abandoning the prior assumption of maximizing the minimum interval successfully obtained in the support vector machine; introducing the optimization goal of interval distribution; and transforming from highlighting the minimum interval formed by specific sample points to highlighting the distribution of the overall data features, thereby weakening the influence of noise data on training and improving the generalization ability of the model [20].

Furthermore, the interval distribution logistic regression model is extended to the task of multiclassification, and by introducing structured regularization, the defect of independent classifiers in the traditional multiclassifier construction is improved, and the shared information in multiple categories is used to improve the overall effect of the model. At the same time, a general framework in the linear case of interval distribution logistic regression is derived. Under this general framework, tasks such as binary classification, multiclassification, and feature selection can be completed at the same time [21]. Moreover, since there are a lot of linear inseparable data in real data, how to use interval distribution logistic regression to construct nonlinear classifiers is also an important content in interval distribution logistic regression research. In this article, by introducing the kernel method, due to interval distribution logistic regression, the simplicity of the model itself can easily be extended to the logistic regression of the kernel interval distribution and successfully used in the scenario of nonlinear classification.

Robustness is an important attribute of a classification model. Robustness is mainly reflected in the ability of the model to adapt to outliers. This ability mainly comes from the definition of the loss function. An ideal loss function should exist in the structure in Figure 1. But the ideal loss function is a nonconvex loss function, which will bring optimization difficulties. Therefore, an ideal and reasonable robust loss function needs to have monotonically decreasing properties and be insensitive to outliers; that is, for classification, the error gives linear loss growth.

Different loss functions are discussed. The squared loss or exponential loss is much more sensitive to outliers than the hinge loss or logistic loss, and it will impose penalties on correctly classified sample points. Among them, the hinge loss is adopted by the support vector machine, and the logistic loss is adopted by the logistic regression. However, since the Hinge Loss is a non-smooth loss term, it may introduce some complexity to the model optimization. Therefore, logic loss is a better alternative. A lot of work focuses on the ability to further explore the logic loss. Vapnik compared logistic regression and support vector machines and proved that chain loss can be approximated by logistic loss. Furthermore, Zhang and Ole proposed generalized logic loss (GLL). The generalized logic loss can approximate the soft-margin support vector machine (Soft-Margin SVM) well under certain conditions, which shows that the logic loss is very important for constructing a simple and robust classifier [22].

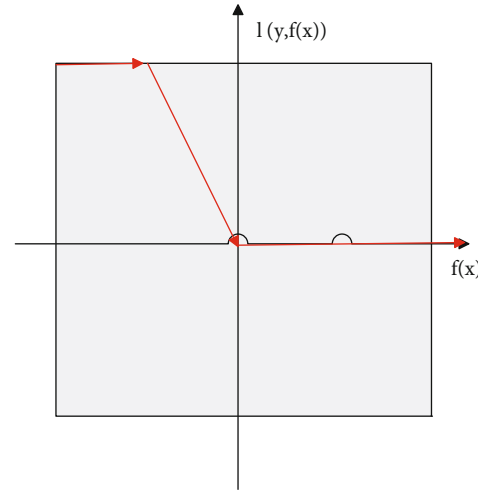


Figure 1: Ideal classification loss function.

When $x \in R^{D\times 1}$ represents a data point and $y \in \{-1,+1\}$ represents the corresponding binary label, the logistic regression model can be expressed as

$$\Pr(y|x) = \frac{1}{1+\exp(-yw^T x)}. \tag{2}$$

Among them, $\Pr(y|x)$ represents the conditional probability of label y in a given sample x, and $w^T x = 0$ defines a classification hyperplane in the feature space. The category conditional probabilities of data points on this classification hyperplane are all 0.5. It should be noted that these data points need to be centralized. Otherwise, x and w should be augmented to [23]

$$\begin{aligned} x &= [x_1, x_2, \cdots, x_D, 1]^T \in R^{(D+1)\times 1},. \\ w &= [w_1, w_2, \cdots, w_D, w_0]^T \in R^{(D+1)\times 1}. \end{aligned} \tag{3}$$

In order to optimize this logistic regression model, we need to optimize the log likelihood function for w:

$$\mathrm{NLL}(w) = -\mathrm{In}\prod_{i=1}^{N}\Pr(y_i|x_i) = \sum_{i=1}^{N}\mathrm{In}\left(1+\exp\left(-y_i w^T x_i\right)\right). \tag{4}$$

The above formula represents the logic loss. The generalized logistic regression loss is proposed to approximate the chain loss used in support vector machines:

$$\mathrm{GLL}\left(\alpha, yw^T x\right) = \frac{1}{\alpha}\mathrm{In}\left(1+\exp\left(-\alpha\left(yw^T x-1\right)\right)\right). \tag{5}$$

The main difference between GLL and logistic loss is that GLL defines the function margin between the two categories. The optimization objective function of GLL is

defined as [24]

$$w = \arg\min_{w} \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\alpha}\mathrm{In}\left(1+\exp\left(-\alpha\left(y_i w^T x_i - 1\right)\right)\right)$$
$$= \arg\min_{w} \frac{1}{N\alpha} 1_N^T \mathrm{In}\left(1_N + \exp\left(-\alpha\left(Y \odot X^T w - 1_N\right)\right)\right). \tag{6}$$

The generalized logic loss is a loss function that can be adjusted by the parameter. When α increases, the generalized logic loss can approximate the chain loss very well. As shown in Figure 2, when $\alpha = 10$, the generalized logic loss is almost exactly the same as the chain loss. Since chain loss is a nonsmooth loss function and logic loss does not impose zero loss on any correctly classified data, which will lead to overlearning of correctly classified samples, generalized logistic loss is a better replacement for loss function.

The success of support vector machines shows that the a priori hypothesis of maximizing the minimum interval can significantly improve the generalization ability of the model. The assumption of maximizing the minimum interval is also applied to logistic regression. Under the framework of generalized logistic loss, the chain loss in support vector machine is replaced with generalized logistic loss, and the maximum interval logistic regression is proposed to approximate support vector machine. Furthermore, the maximum interval logistic regression is also extended to feature selection and sparse learning. But even if the maximum interval strategy is effective most of the time, this strategy is easily affected on noisy data.

The interval distribution, by considering the interval distribution of all data points, rather than the minimum interval of the data points closest to the decision boundary, is proved to have better performance than the minimum interval strategy.

Studies have shown that all data points have an impact on the generalization error boundary, and the impact of a data point on the generalization error and its distance from the decision boundary show an exponential decrease; that is, the closer the point to the decision boundary, the more impact on the generalization error (big). The traditional support vector machine is proved to have a lower bound on the divergence between data classes, but it ignores the important prior distribution information in the data classes [25].

Theorem 1.

$$\Pr_D[yf(x) < 0] \le \frac{1}{m^{50}} + \inf_{\theta\in(0,1]}\left[\Pr_S[yf(x) < \theta] + m^{-2/\left(1-E_S^2[yf(x)+\theta/9]\right)}\right] + \frac{3\sqrt{\mu}}{m^{3/2}} + \frac{7\mu}{3m} + \sqrt{\frac{3\mu}{m}\widehat{L}(\theta)}. \tag{7}$$

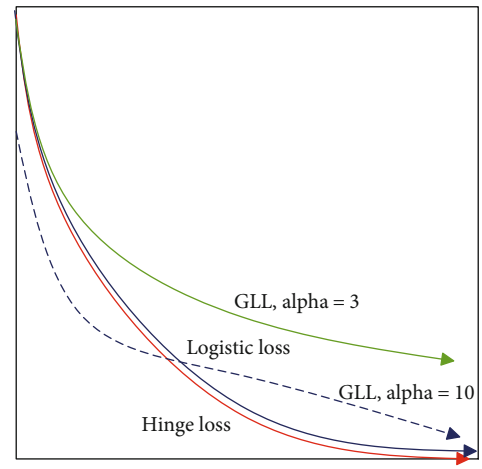


Figure 2: Comparison of different loss functions.

Among them,

$$\mu = \frac{144InmIn(2|H|)}{\theta^2} + \mathrm{In}\left(\frac{2|H|}{\delta}\right),$$
$$\widehat{L}(\theta) = \Pr_S[yf(x) < \theta]\Pr_S\left[yf(x) > \frac{2\theta}{3}\right]. \tag{8}$$

$Es[yf(x)]$ is the mean of the margin, and $\widehat{L}(\theta)$ is the variance of the margin.

Theorem 1 proves that the margin mean and margin variance play a key role in the generalization of the classifier. The margin distribution is defined as the first and second moments of the margin, that is, the margin mean $(\bar{\eta})$ and the margin variance $(\widehat{\eta})$. The optimization goal of the margin distribution is to simultaneously maximize the margin mean and minimize the margin variance:

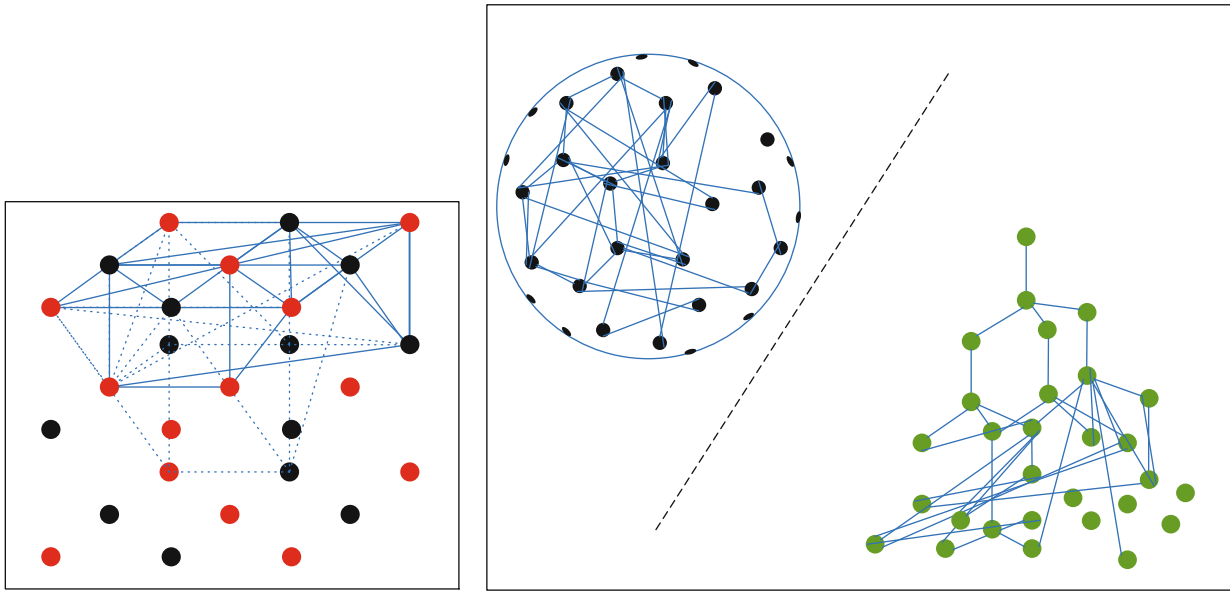
$$\bar{\eta} = \frac{1}{N}\sum_{i=1}^{N} y_i w^T x = \frac{1}{N}(XY)^T w, \tag{9}$$

$$\widehat{\eta} = \frac{1}{N}\sum_{i=1}^{N}\left(y_i w^T x_i - \bar{\eta}\right)^2. \tag{10}$$

The margin distribution has a significant impact on the learned classification hyperplane. It will help the classification model to fully consider the statistical information hidden in the training data:

Theorem 2. *The margin mean value in formula (9) will help expand the class center distance of the two types, and the margin variance constraint in formula (10) will force the hyperplane to be in a direction with higher data uncertainty and prevent the hyperplane from deviating too much from the centers of the two types.*

3.1. Certification. N_+ and N_- are the numbers of positive and negative samples, S_+ and S_- are the sets of positive and negative samples, respectively, $\bar{x}_{+--}$ and $\bar{x}_-$ are the centers of positive and negative samples, respectively, and S_W^+ and S_W^-



(a) Hyperplanes in feature space (b) Projected points in logistic space

Figure 3: Comparison of logistic regression and margin distribution logistic regression.

are the covariance matrices of positive samples and negative samples, respectively. The mean margin can be correspondingly expressed as

$$\begin{aligned}\bar{\eta} &= \frac{1}{N}\sum_{i=1}^{N} y_i w^T x_i = \frac{1}{N}\left(\sum_{x_i \in S_+} w^T x_i - \sum_{x_i \in S_-} w^T x_i\right) \\ &= \frac{N_+}{N} w^T \bar{x}_{+--} - \frac{N_-}{N} w^T \bar{x}_{---}.\end{aligned} \tag{11}$$

Similarly, the margin distribution can be rewritten as

$$\begin{aligned}\hat{\eta} &= \frac{1}{N}\sum_{i=1}^{N} \left(y_i w^T x_i - \bar{\eta}\right)^2 \\ &= \frac{1}{N}\left(\sum_{x_i \in S_+} \left(w^T x_i - \bar{\eta}\right)^2 - \sum_{x_i \in S_-} \left(w^T x_i + \bar{\eta}\right)^2\right).\end{aligned} \tag{12}$$

Part of the above formula is expressed as

$$\begin{aligned}\sum_{x_i \in S_+} \left(w^T x_i - \bar{\eta}\right)^2 &= \sum_{x_i \in S_+} \left(w^T x_i - \left(\frac{N_+}{N} w^T \bar{x}_{+--} - \frac{N_-}{N} w^T \bar{x}_{---}\right)\right)^2 \\ &= N_+ w^T S_W^+ w + \frac{N_+^2 N_-}{N^2} w^T \left(\bar{x}_+ + \bar{x}_{----}\right)^2 w.\end{aligned} \tag{13}$$

Thus, the margin variance can be further rewritten as

$$\hat{\eta} = \frac{1}{N} w^T \left(N_+ S_W^+ + N_- S_W^-\right) w + \frac{N_+ N_-}{N^2} w^T \left(\bar{x}_{+--} + \bar{x}_-\right) w. \tag{14}$$

For the margin mean formula, when the margin mean is optimized, the distance between the two categories is enlarged, thereby improving the discriminative ability of the model. For the margin variance, it can be decomposed into two parts. The first part represents the margin variance of each category. When the margin variance is minimized, the discriminant hyperplane will be along the direction with the greatest data uncertainty, thereby reducing the possibility of data crossing the discriminant hyperplane. When the second part of the margin variance is maximized, the discriminant hyperplane will not deviate too much from the two types of center points, thereby obtaining a more reasonable discriminant hyperplane and ensuring the discriminative ability of the model.

Through the analysis of the margin distribution characteristics, it can be known that when adjusting the parameters of the margin distribution constraint, it can help the model to better adapt to the distribution of the training data. When the model can make better use of the statistical information in the data, the model can have better adaptability to noise and outliers.

A margin distribution logistic regression model with robustness and generalization ability is defined as

$$\arg\min_{w} \frac{1}{N\alpha} 1_N^T \mathrm{In}\left[1_N + \exp\left(-\alpha\left(Y \odot X^T w - 1_N\right)\right)\right] + \lambda_1 \hat{\eta} - \lambda_2 \bar{\eta}. \tag{15}$$

In the above formula, the classification error is minimized by GLL, while the margin variance is reduced and the margin mean is increased. The margin is the functional distance of a sample point to distinguish the hyperplane. However, after the logistic regression projection, the margin corresponds to the distance of the $w^T x$ from the origin of the coordinate after the data point is projected. Among them, the origin of the coordinates represents the position where the classification probability is 0.5, and the slope of the curve

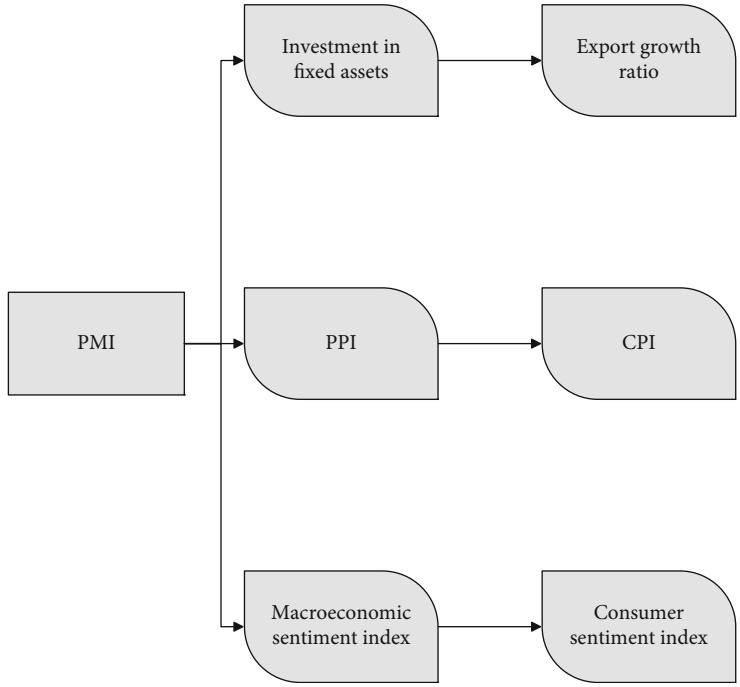


Figure 4: Mechanism analysis of PMI and different indicators.

here is the highest. Obviously, data points near the origin are more likely to be classified incorrectly. When we increase the mean margin, as shown in Figure 3(b), we can make a larger safety margin between the two categories and make the classification confidence higher. Moreover, the margin distribution also makes the corresponding discriminant hyperplane changes to the data distribution, as shown in Figure 3(a). When the control margin variance is small, the discrimination hyperplane can be along the direction of greater data uncertainty. At the same time, due to the optimization of the margin mean, the learned classification hyperplane can still maintain its most critical discriminative ability in the classification task.

In real applications, most data is composed of multiple categories. Therefore, it is of great significance to extend the two classifiers to multiple classifiers. In this paper, we extend the margin distribution logistic machine (MDLM) to a multiclass version.

Generally speaking, a multiclass classifier determines the final category by combining a group of one-to-many two-classifiers or a group of one-to-one two-classifiers through voting. However, these two combination strategies have a common flaw. That is, there is no way for these independent two classifiers to share information, resulting in the information between categories cannot be shared. For example, there is a feature subset that is effective for this multiclassification task, but an independent set of classifiers cannot help capture this important information. Therefore, we need to adopt a method that can not only classify at the same time but also help the model capture the information that exists between the categories. There are also studies that combine a C-class classification task into an optimization and avoid the complexity of training a large number of independent

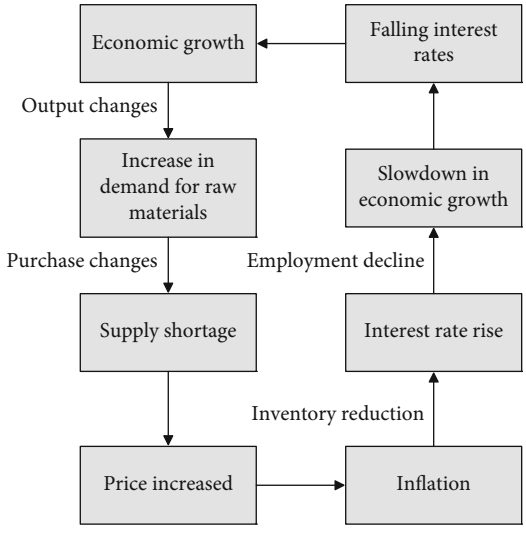


Figure 5: The connection between the PMI system and the business cycle.

classifiers at the coding and logic levels. Through the following method, a multiclass support vector machine is constructed, and this multiclass support vector machine can be optimized by a model:

$$\begin{aligned} \min \quad & \left\{ C \sum_{i=1}^{n} \sum_{k \neq y_i} \xi_i^k + \frac{1}{2} \sum_{k=1}^{C} \sum_{j=1}^{m} \left(w_{k,j} \right)^2 \right\} \\ \text{s.t.} \quad & w_{y_i}^T x_i \geq w_k^T x_i + 2 - \xi_i^k \\ & \xi_i^k \geq 0 \quad (i = 1, 2, \cdots, n\,; k \neq y_i). \end{aligned} \tag{16}$$

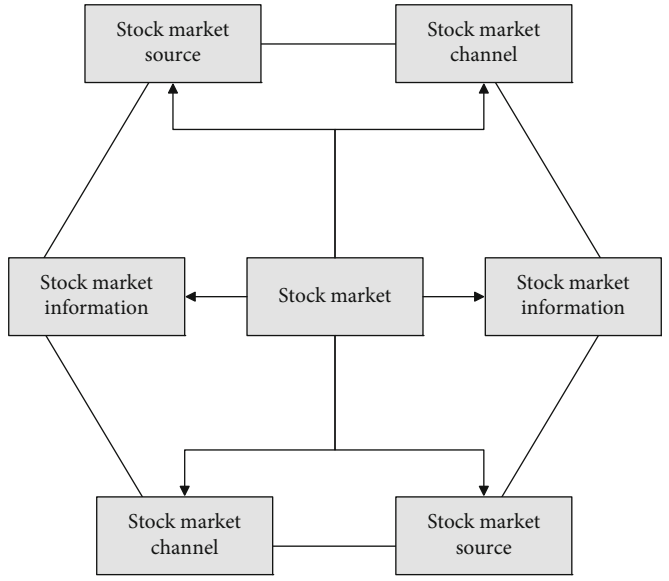


Figure 6: The information transmission model of the stock market.

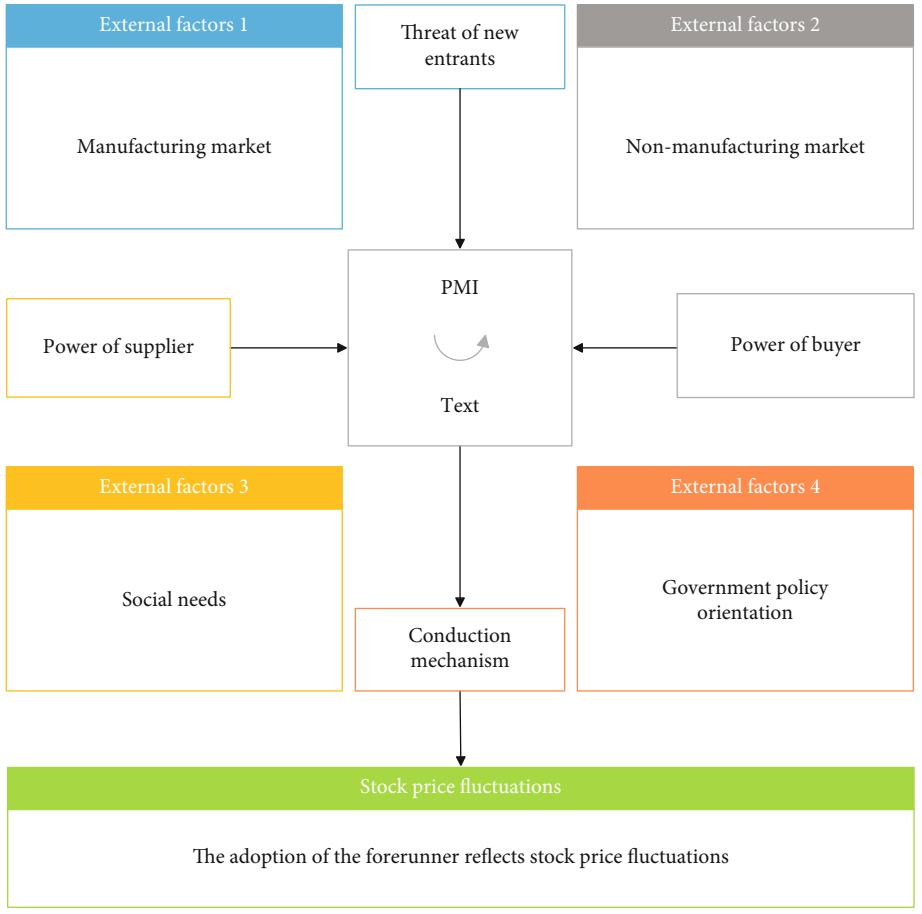


Figure 7: Correlation analysis model of stocks and PMI index based on logistic regression model.

GLL is used to replace the nonsmooth chain loss in support vector machines. Similarly, the SVM using GLL has also been extended to a multiclass model. This multi-class maximum margin logistic regression (MLR) is given in the following form. This mode can also complete the learning of a multiclass classifier in an optimization

Table 1: Statistical table of China's PMI from December 2017 to March 2021.

	Manufacturing PMI	Year-on-year growth	Nonmanufacturing PMI	Year-on-year growth
1-Mar-21	51.9	-0.19%	56.3	7.65%
1-Feb-21	50.6	41.74%	51.4	73.65%
1-Jan-21	51.3	2.60%	52.4	-3.14%
1-Dec-20	51.9	3.39%	55.7	4.11%
1-Nov-20	52.1	3.78%	56.4	3.68%
1-Oct-20	51.4	4.26%	56.2	6.44%
1-Sep-20	51.5	3.41%	55.9	4.10%
1-Aug-20	51	3.03%	55.2	2.60%
1-Jul-20	51.1	2.82%	54.2	0.93%
1-Jun-20	50.9	3.04%	54.4	0.37%
1-May-20	50.6	2.43%	53.6	-1.29%
1-Apr-20	50.8	1.40%	53.2	-2.03%
1-Mar-20	52	2.97%	52.3	-4.56%
1-Feb-20	35.7	-27.44%	29.6	-45.49%
1-Jan-20	50	1.01%	54.1	-1.10%
1-Dec-19	50.2	1.62%	53.5	-0.56%
1-Nov-19	50.2	0.40%	54.4	1.87%
1-Oct-19	49.3	-1.79%	52.8	-2.04%
1-Sep-19	49.8	-1.97%	53.7	-2.19%
1-Aug-19	49.5	-3.51%	53.8	-0.74%
1-Jul-19	49.7	-2.93%	53.7	-0.56%
1-Jun-19	49.4	-4.08%	54.2	-1.45%
1-May-19	49.4	-4.82%	54.3	-1.09%
1-Apr-19	50.1	-2.53%	54.3	-0.91%
1-Mar-19	50.5	-1.94%	54.8	0.37%
1-Feb-19	49.2	-2.19%	54.3	-0.18%
1-Jan-19	49.5	-3.51%	54.7	-1.08%
1-Dec-18	49.4	-4.26%	53.8	-2.18%
1-Nov-18	50	-3.47%	53.4	-2.55%
1-Oct-18	50.2	-2.71%	53.9	-0.74%
1-Sep-18	50.8	-3.05%	54.9	-0.90%
1-Aug-18	51.3	-0.77%	54.2	1.50%
1-Jul-18	51.2	-0.39%	54	-0.92%
1-Jun-18	51.5	-0.39%	55	0.18%
1-May-18	51.9	1.37%	54.9	0.73%
1-Apr-18	51.4	0.39%	54.8	1.48%
1-Mar-18	51.5	-0.58%	54.6	-0.91%
1-Feb-18	50.3	-2.52%	54.4	0.37%
1-Jan-18	51.3	0.00%	55.3	1.28%
1-Dec-17	51.6	0.39%	55	0.92%

problem:

$$\min \frac{1}{n\alpha}\sum_{i=1}^{n}\sum_{k\neq y_i}\mathrm{In}\left(1+\exp\left(-\alpha\left(w_{y_i}^{T}-w_k^{T}\right)X-2\right)\right) + \lambda\sum_{k=1}^{C}\sum_{j=1}^{m}\left(w_{k,j}\right)^2. \tag{17}$$

In this paper, using the core idea of multiclass classifiers, an equivalent multiclass learning model is given. We assume that a given data set has different categories of C and construct a multiclass label matrix $Y=[y_1,y_2,\cdots,y_C]\in\{-1,1\}^{N\times C}$ of this data set and $y_i=[-1,\cdots-1,\underbrace{1,\cdots,1}_{i-\text{thclass}},-1,\cdots-1]^T$. This multiclass classifier is

built on the basis of MDLM. Its optimization objective function is

$$\arg\min w = [w_1,\cdots,w_C]\sum_{i=1}^{C}\frac{1}{N\alpha}1_N^T\mathrm{In}\left[1_N+\exp\left(-\alpha\left(y_i\odot X^T w_i-1_N\right)\right)\right] + \lambda_1\widehat{\eta}_i-\lambda_2\bar{\eta}_i+\beta\|w\|_{2,1}. \tag{18}$$

This multiclass MDLM model tries to learn C one-to-many multiclassifiers at the same time. It is worth noting that it is different from the L2 norm regularity imposed in SVM and large margin logistic regression; in this paper, we use the L21 norm to capture the information implicit in each subtask in this multiclass task, especially the feature validity information. In particular, when the number of categories C is equal to 2, the model can be adjusted to the L1 norm, that is, the two-class sparse margin distribution logistic machine model (SMDLM). The L1 norm has also been proven to be a very effective technique for classification problems, and the SMDLM model can be well applied in many feature selection problems. When the parameter β is equal to 0, the model degenerates to MDLM. Therefore, such a multiclass MDLM model can be well applied in the expansion of two classification, feature selection, multiclassification, multimodel learning, multimodal feature selection, and so on.

In real applications, nonlinear classification is also a very important research content, because there are a large number of linear inseparable data sets.

Research on the nonlinear model of logistic regression has also received a lot of attention. The objective function of linear margin distribution logistic regression is

$$\arg\min_{w}\frac{1}{N\alpha}1_N^T\mathrm{In}\left[1_N+\exp\left(-\alpha\left(Y\odot X^T w-1_N\right)\right)\right]+\lambda_1\widehat{\eta}-\lambda_2\bar{\eta}. \tag{19}$$

Any function can be expressed in the following form:

$$f(x)=\sum_{i=1}^{n}\alpha_i K(x,x_i). \tag{20}$$

Among them, $K(x,x_i)$ represents the kernel function, which can use linear kernel function, polynomial kernel function, radial basis kernel function, and so on. This paper defines K as the radial basis kernel function:

$$K(x,x_i)=e^{-\left(\|x-x_i\|^2\right)/2\sigma^2}. \tag{21}$$

If we define K as the sample kernel matrix, $K\in R^{N\times N}$, then,

$$f(x)=w^T x=\sum_{i=1}^{n}\alpha_i K(x,x_i)=\alpha^T K_i. \tag{22}$$

The objective function of the nonlinear margin distribution logistic regression is obtained:

$$\arg\min_{\alpha}\frac{1}{N\alpha}1_N^T\mathrm{In}\left[1_N+\exp\left(-\alpha\left(Y\odot K^T\alpha-1_N\right)\right)\right] - \frac{\lambda_2}{N}Y^T K^T\alpha+\frac{\lambda_1}{N}K^T\alpha\alpha^T K-\frac{\lambda_1}{N^2}\alpha^T KYY^T K^T\alpha. \tag{23}$$

The nonlinear margin distribution logistic regression algorithm is a smooth and convex model, which can be directly optimized by the gradient descent algorithm.

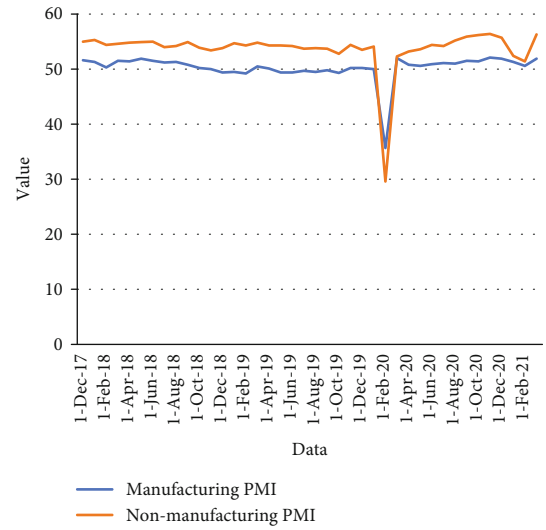


Figure 8: Statistical diagram of China's PMI from December 2017 to March 2021.

4. Analysis Model of the Correlation between Stocks and PMI Index Based on Logistic Regression Model

PMI has a linkage relationship with the early warning index of the macroeconomic prosperity index, the consumer expectations index, and the real estate prosperity index. PMI is also an indicator of economic prosperity. Therefore, this paper empirically analyzes the correlation between PMI and other economic prosperity indexes, which are also economic prosperity indexes. PMI has the leading value in the economic climate index and reflects the characteristics of different economic cycles of economic fluctuations, as shown in Figure 4.

The mechanism of PMI's effect on the macro economy is shown in Figure 5.

According to the theory of supply and demand, the prices of labor and raw materials will rise, and therefore, the price level of the society as a whole will rise, and the economy as a whole is in a state of inflation. At this time, the country will take control of the supply of currency to ensure the stability of the price level, and the reduction of the supply of money will make it difficult for companies to meet the demand for funds, and interest rates will rise. The increase in interest rates will increase the cost of inventories of enterprises, reduce inventories, and relatively decrease

Table 2: Statistical table of the correlation between stock price and PMI from December 2017 to March 2021.

Data	Manufacturing PMI	Share price	Data	Manufacturing PMI	Share price
1-Mar-21	51.9	13.8	1-Jul-19	49.7	15.3
1-Feb-21	50.6	15.9	1-Jun-19	49.4	14.9
1-Jan-21	51.3	17.1	1-May-19	49.4	14.2
1-Dec-20	51.9	15.5	1-Apr-19	50.1	15.5
1-Nov-20	52.1	17.4	1-Mar-19	50.5	15.2
1-Oct-20	51.4	16.6	1-Feb-19	49.2	16.6
1-Sep-20	51.5	14.7	1-Jan-19	49.5	16.3
1-Aug-20	51	16.9	1-Dec-18	49.4	14.3
1-Jul-20	51.1	15.6	1-Nov-18	50	15.7
1-Jun-20	50.9	16.2	1-Oct-18	50.2	16.5
1-May-20	50.6	16.4	1-Sep-18	50.8	15.1
1-Apr-20	50.8	8.6	1-Aug-18	51.3	16.0
1-Mar-20	52	15.2	1-Jul-18	51.2	16.2
1-Feb-20	35.7	15.6	1-Jun-18	51.5	16.3
1-Jan-20	50	16.9	1-May-18	51.9	15.6
1-Dec-19	50.2	14.4	1-Apr-18	51.4	14.7
1-Nov-19	50.2	15.2	1-Mar-18	51.5	15.3
1-Oct-19	49.3	15.3	1-Feb-18	50.3	15.8
1-Sep-19	49.8	14.5	1-Jan-18	51.3	15.1
1-Aug-19	49.5	14.4	1-Dec-17	51.6	13.7

production. On the other hand, the increase in interest rates will also slow down the momentum of overall investment and consumption. In turn, our country's economy, which currently mainly relies on these two methods to drive economic growth, has gradually slowed down its growth rate.

The manufacturing PMI data is authoritatively published by the China Logistics Purchasing Federation, while the HSBC Manufacturing PMI is published by HSBC. The channels of the stock market mainly include relevant channels for the transmission of economic information, such as the Internet, newspapers, research reports, mobile phones, and other media. The information sink of the stock market is the investors of all kinds of stock markets. Investors use their own analysis and judgment of economic information and make corresponding stock trading behaviors, thereby affecting the fluctuation or trend of the stock market. Noise in the stock market refers to the fact that macroeconomic information is inevitably affected by other external factors in the process of transmission, including both human and nonhuman factors. Under the combined influence of these factors, macroeconomic information will be disturbed by noise such as exaggeration, reduction, and distortion, which will affect the transmission of information and ultimately affect investors' investment decisions. The information transmission model of the stock market is shown in Figure 6.

Combining the demand for stock price fluctuation analysis, the correlation analysis model of stock and PMI index based on logistic regression model constructed in this paper is shown in Figure 7.

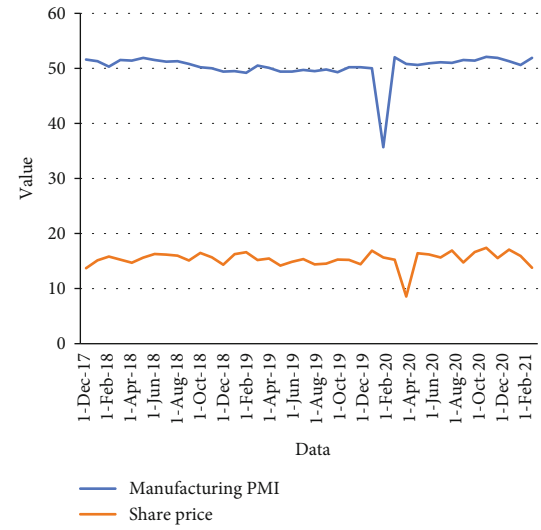


Figure 9: Statistical diagram of the correlation between stock price and PMI from December 2017 to March 2021.

5. Test Analysis

According to actual needs, a correlation analysis model of stocks and PMI based on logistic regression model is constructed. On this basis, this paper obtains data from the National Bureau of Statistics to analyze the performance of the model and count the PMI data of China in recent years, as shown in Table 1 and Figure 8.

This article takes the stock price of a listed company as the evaluation object and analyzes the correlation with PMI. The research object of this paper is manufacturing

enterprises, so the manufacturing PMI is selected as the research object. The statistical results are shown in Table 2 and Figure 9.

Through the analysis of the above chart and table, we can see that the stock price has a strong correlation with the PMI. In addition, compared with stock prices, PMI is at least two months forward-looking. Therefore, the system model constructed in this paper shows that there is a clear correlation between PMI and stocks.

6. Conclusion

Based on the reality that the Purchasing Managers Index is widely used in the forecast of economic growth and is known as the "barometer" of the stock market, this paper first theoretically analyzes the forecasting and functioning mechanism of the Purchasing Manager Index for the macroeconomy and the stock market. Moreover, this paper designs a margin distribution logistic regression model that is easy to optimize and has robustness and generalization ability on the basis of generalized logistic loss and margin distribution and gives a multiclass margin distribution logistic regression framework. This framework can be used to perform two-classification, multiclassification, and feature selection tasks. In addition, this paper constructs an analysis model of the correlation between stocks and the PMI index based on a logistic regression model and combines actual data to carry out the analysis. From the experimental results, it can be seen that the algorithm model constructed in this paper has certain practical effects.

Data Availability

The labeled datasets used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare no competing interests.

Acknowledgments

This study is sponsored by Henan Industrial Vocational and Technical College.

References

[1] C. J. Adam, "How firms shape income inequality: stakeholder power, executive decision making, and the structuring of employment relationships," *Academy of Management Review*, vol. 41, no. 2, pp. 324–348, 2016.

[2] M. B. Alexandri and W. K. Anjani, "Income smoothing: impact factors, evidence in Indonesia," *International Journal of Small Business and Entrepreneurship Research*, vol. 3, no. 1, pp. 21–27, 2014.

[3] T. Alquraan, A. Alqisie, and S. A. Al, "Do behavioral finance factors influence stock investment decisions of individual investors?(Evidences from Saudi Stock Market)," *Journal of American Science*, vol. 12, no. 9, pp. 72–82, 2016.

[4] N. Apergis, B. Simo-Kengne, and R. Gupta, "The long-run relationship between consumption, house prices, and stock prices in South Africa: evidence from provincial-level data," *Journal of Real Estate Literature*, vol. 22, no. 1, pp. 83–99, 2014.

[5] N. Arshed, A. Anwar, M. S. Hassan, and S. Bukhari, "Education stock and its implication for income inequality: the case of Asian economies," *Review of Development Economics*, vol. 23, no. 2, pp. 1050–1066, 2019.

[6] S. R. Baker, "Debt and the response to household income shocks: validation and application of linked financial account data," *Journal of Political Economy*, vol. 126, no. 4, pp. 1504–1557, 2018.

[7] M. Banam and A. Mehrazeen, "The relationship of information asymmetry, institutional ownership and stock liquidity with income smoothing in Tehran Stock Exchange," *Journal OF Management and Accounting Studies*, vol. 4, no. 3, pp. 6–11, 2016.

[8] E. Bengtsson and D. Waldenström, "Capital shares and income inequality: evidence from the long run," *The Journal of Economic History*, vol. 78, no. 3, pp. 712–743, 2018.

[9] L. Carvalho and C. Di Guilmi, "Technological unemployment and income inequality: a stock-flow consistent agent-based approach," *Journal of Evolutionary Economics*, vol. 30, no. 1, pp. 39–73, 2020.

[10] S. Chen and B. Komal, "Impact of stock market development on economic growth: evidence from lower middle income countries," *Management and Administrative Sciences Review*, vol. 5, no. 2, pp. 86–97, 2016.

[11] Y. Dafermos and C. Papatheodorou, "Linking functional with personal income distribution: a stock-flow consistent approach," *International Review of Applied Economics*, vol. 29, no. 6, pp. 787–815, 2015.

[12] R. Gantino, "Effect of managerial ownership structure, financial risk and its value on income smoothing in the Automotive Industry and Food & Beverage Industry Listed in Indonesia Stock Exchange," *Research Journal of Finance and Accounting*, vol. 6, no. 4, pp. 48–56, 2015.

[13] M. Gao, J. Meng, and L. Zhao, "Income and social communication: the demographics of stock market participation," *The World Economy*, vol. 42, no. 7, pp. 2244–2277, 2019.

[14] J. P. Gómez, R. Priestley, and F. Zapatero, "Labor income, relative wealth concerns, and the cross section of stock returns," *Journal of Financial and Quantitative Analysis*, vol. 51, no. 4, pp. 1111–1133, 2016.

[15] M. Hoffmann, "The consumption–income ratio, entrepreneurial risk, and the U.S. stock Market," *Journal of Money, Credit and Banking*, vol. 46, no. 6, pp. 1259–1292, 2014.

[16] B. F. Jones, "The human capital stock: a generalized approach," *American Economic Review*, vol. 104, no. 11, pp. 3752–3777, 2014.

[17] H. H. Khan, I. Naz, F. Qureshi, and A. Ghafoor, "Heuristics and stock buying decision: evidence from Malaysian and Pakistani stock markets," *Borsa Istanbul Review*, vol. 17, no. 2, pp. 97–110, 2017.

[18] H. J. Kleven and E. A. Schultz, "Estimating taxable income responses using Danish tax reforms," *American Economic Journal: Economic Policy*, vol. 6, no. 4, pp. 271–301, 2014.

[19] G. Li, "Information sharing and stock market participation: evidence from extended families," *Review of Economics and Statistics*, vol. 96, no. 1, pp. 151–160, 2014.

[20] M. Y. Mohammadi and M. H. Arman, "The survey of accounting variables effect on incomesmoothing in stock exchange companies," *Journal of Fundamental and Applied Sciences*, vol. 8, no. 2, pp. 1257–1271, 2016.

[21] S. N. Nazar, T. Ekowati, and H. Setiyawan, "Does income smoothing improve informativeness of stock prices?," *Econosains Jurnal Online Ekonomi dan Pendidikan*, vol. 15, no. 2, pp. 225–239, 2017.

[22] T. Nguyen, H. N. Duong, and H. Singh, "Stock market liquidity and firm value: an empirical examination of the Australian market," *International Review of Finance*, vol. 16, no. 4, pp. 639–646, 2016.

[23] E. Saez and G. Zucman, "Wealth inequality in the United States since 1913: evidence from capitalized income tax data," *The Quarterly Journal of Economics*, vol. 131, no. 2, pp. 519–578, 2016.

[24] S. H. Shin and K. T. Kim, "Income uncertainty and household stock ownership during the great recession," *Journal of Financial Counseling and Planning*, vol. 29, no. 2, pp. 383–395, 2018.

[25] J. Voelzke, "Individual labour income, stock prices and whom it may concern," *Applied Economics Letters*, vol. 23, no. 13, pp. 965–968, 2016.