International Journal for Housing Science and Its Applications
Publish August 3, 2025. Volume 46, Issue 3 Pages 805-815

https://doi.org/10.70517/ijhsa46360

Research on Constructing Enterprise Data Asset Pricing Model with the Help of Principal Component Analysis in the Era of Digital Economy

Daohan Zhang[1,*] and Feng Li[2,3]
[1] School of Management, Liaodong University, Dandong, Liaoning, 118001, China
[2] School of Accounting, Luoyang Institute of Science and Technology, Luoyang, Henan, 471023, China
[3] School of Business, Silla University, Busan, 617736, South Korea

Corresponding authors: (e-mail: bixiaqiu1991@163.com).

Abstract In this paper, we mined the financial data of several enterprises in CSMAR database and constructed an asset pricing model based on level, slope, and curvature (LSC). By combing the characteristics of data assets and their value drivers, the key factors affecting asset pricing are extracted with the help of principal component analysis. Then the relationship between corporate credit risk and asset excess return is analyzed by using Fama-Macbeth one-factor regression method. And combined with GRS and other tests, the explanatory power of asset pricing on returns under different models is comparatively assessed. The credit risk coefficients of the models in the Fama-Macbeth test range from -0.473 to -0.115, and there is a significant negative correlation between them and the portfolio excess returns. There are monotonically increasing or decreasing trends in indicators such as expected excess return and average excess return. And the predicted return of the asset pricing method in this paper is closer to the real return, with the same upward trend. The first three components obtained by principal component analysis explain 90.3%, 3.7%, and 1.4% of the portfolio, respectively. The GRS statistics of the LSC pricing model in this paper are lower than the baseline model by 0.286 to 0.930, which has stronger pricing explanatory power. This study expands the theoretical framework of data asset pricing, which is of value in the marketized allocation of enterprise data assets in the era of digital economy.

Index Terms Asset pricing model, Principal component analysis, Fama-Macbeth regression, GRS test

I. Introduction

Currently, China's digital economy is in a phase of rapid growth. According to the Digital China Development Report (2022), the size of China's digital economy has exceeded the 50 trillion yuan mark in 2022 and accounts for 41.5% of GDP, ranking second globally [1]. This achievement highlights China's remarkable progress in digital transformation and digital economy development. With the booming digital industry, experts have begun to work on exploring how to use data-enabled tools to drive social progress. However, the issue of effective management and utilization of massive data in this process appears relatively neglected, failing to fully explore the potential value contained in these data resources [2], [3].

Reliable data asset pricing can help to make up for the shortcomings of the current data resource accounting system, promote the release of the value of data resources, and more systematically and scientifically evaluate the contribution of data elements to economic and social development [4]. However, enterprise data asset pricing is a challenging task. On the one hand, compared with other traditional tangible and intangible assets, data assets have differentiated characteristics such as non-consumption, non-competition, non-exclusivity, economies of scale, and zero-cost replication [5], [6]. On the other hand, there are many factors affecting the value of data assets, and the data's own characteristics, application scenarios, etc., will have an important impact on the value of data assets, and how to portray these influencing factors is another difficulty facing the assessment of the value of data assets [7].

The traditional asset pricing method mainly calculates the price of commodities through cost, and after a comprehensive discussion of data cost in the research of Goldfarb, A and Tucker, C et al, it is found that the data asset has high sunk cost and marginal cost close to zero, and the traditional method can not be applied to data pricing [8]. Therefore, domestic and foreign scholars have launched a new round of research on data pricing methods. Discussing the problem of data asset pricing, it is necessary to clarify which data can be involved in pricing as a production factor in the first place, Pei, J pointed out in his research that not all data assets can be priced, and only data products such as datasets and data packages are the real objects of data asset pricing [9]. According to

different pricing strategies and ideas, this paper will describe the theoretical research results of three types of data pricing methods based on utility, based on information entropy, based on auction and game theory.

The core of data utility-based pricing is to set a price for data assets according to the data's own attributes and utility [10]. Data quality is an important issue, Heckman, J. R et al. focusing on the impact of data quality and data products on the data market, proposed a simple linear dataset-based asset valuation model in which the price of data assets is mainly affected by factors such as data volume, data accuracy and data quality [11]. Yu, H and Zhang, M comprehensively considered multi-version pricing consisting of multiple data quality factors and developed a two-tier data asset pricing model to maximize the profit of data asset owners and maximize the utility of consumers, but the model involves a two-tier planning problem [12]. The data quality dimension is an issue that must be considered for utility-based data pricing, but how data quality is quantified, how to achieve load balance between different dimensions, and how to set the factor weights to be reasonable are all barriers to the ground-level implementation of this pricing method.

The concept of information entropy was proposed by Shannon in 1948, and at the same time, he also pointed out the method of calculating entropy, which was proposed to solve the problem of quantitative measure of information [13]. In data asset pricing, information entropy can describe the uncertainty of useful information within the data [14]. Shen, Y et al. calculated the data price as the entropy of each data tuple based on the value of information entropy, and this pricing model can provide reasonable pricing for personal data based on the value of information entropy [15]. Li, X et al. proposed a pricing function based on the result of data information entropy, verified the reasonableness of the data information measurement method, and gave three specific pricing function models, but this method also can not realize dynamic pricing [16]. In addition to this, data pricing based on information entropy needs to consider the uncertainty of the internal information of the data, which makes pricing more difficult and makes it difficult to accurately measure the data entropy value.

Data asset pricing based on auction and game theory focuses more on the process of setting prices, and this type of data asset pricing method requires the participation of data providers, consumers and data trading platforms [17]. The open sharing of some information in data assets enables buyers and sellers to negotiate prices to promote data turnover [18]. Koutris, P et al. proposed a new linear programming model for query-based data to realize a random sampling auction, this pricing method, although truly reflects the utility of data buyers, but ignores the interests of data sellers, and is difficult to be applied in practice [19]. Xiao, Z et al. proposed a data asset pricing model with three parties (data vendors, service providers, and service users) that uses a Stackelberg game to determine the optimal pricing strategy in order to maximize the profits of all participants [20]. Chen, Y et al. constructed a three-party pricing model for data assets by combining Stackelberg game and auction mechanism, and coordinated buyers, sellers and platforms to ensure fairness through dynamic adjustment mechanism, and achieved 92.7% transaction success rate through simulation validation [21]. Hu-Bolz, J et al. proposed a data asset pricing strategy based on mean-field game theory to model the interactions between owners and consumers in federated data to maximize their mutual interests [22]. The long and complex game process of information exchange is bound to significantly increase the cost of transaction time and reduce the efficiency of data transactions. At the same time, due to information asymmetry, it will make the auction and game theory based data asset pricing methods, the potential actual fair value of itself becomes impossible to accurately measure, leading to the deviation of the value of data assets [23].

Currently, data pricing is dominated by utility-based or auction- and game-theory-based pricing methods, the former of which, while capable of measuring the value of the data to a certain extent, takes a relative pricing approach, determining the relative or virtual price of the data [24]. The latter, although using absolute pricing methods to determine the absolute price of data transactions, but the overall consumption of human and material resources is larger, the asymmetry of information in the intangible also impede the transaction of data, the current relative pricing of data and absolute pricing methods exist in the shortcomings and defects [25]. The above limitations make data trading marketable and have not reached the expected level of activity [26]. Therefore, there is an urgent need to construct a more fair, transparent and efficient data asset pricing model, which is of practical significance to the development and practice of data trading market.

The article collects monthly frequency data from the A-share market in the CSMAR database, and obtains a sample of asset pricing research data through extreme value processing, missing value processing and data normalization. The returns of stocks are predicted through cross-sectional regressions of stocks, and the stocks are equally divided into portfolios based on the expected returns. Then the public pricing factors related to the expected returns are extracted by principal component analysis, and then the asset pricing model of corporate data based on principal component analysis is constructed. Meanwhile, the relationship between corporate credit risk and return is predicted by combining the Fama-Macbeth regression method. And statistics such as average excess return and expected excess return of the portfolio are described to further analyze the asset pricing ability of the model.

II. Data asset pricing analysis techniques

II. A.Principal component analysis

Principal Component Analysis [27], is a statistical method used to explore the relationship between multiple possibly related variables. It decomposes a set of data into different dimensions, and the data on each dimension represents the largest current set of basis vectors. Some principal components reflect the vast majority of information about the original variables and are uncorrelated with each other, making them more intuitive to analyze. The specific process is as follows:

First, the first step requires centering or normalizing a set of spectral data. Suppose a set of samples are $X_1, X_2, X_3, ..., X_n$, the mean of the sample is $\overline{X}$ and its standard deviation is σ.

Then the centering is calculated as:

$$Y_i = X_i - \overline{X} \tag{1}$$

The standardized calculation method is:

$$Y_i = \frac{X_i - \overline{X}}{\sigma} \tag{2}$$

In the second step, the covariance matrix of the data matrix is computed. If there are a total of n sets of data after centering and each set contains m variables. Then the data can be viewed as a matrix of order $n \times m$, i.e.:

$$X_i = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1m} \\ X_{21} & X_{22} & \cdots & X_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{nm} \end{bmatrix} \tag{3}$$

Then its covariance matrix is:

$$D(X) = \operatorname{cov}(X_i, X_j) i, j = 1, 2, 3, ..., n \tag{4}$$

Also writing:

$$D(X) = \begin{bmatrix} E[(X_l - E(X_l))(X_l - E(X_l))] & \cdots & E[(X_l - E(X_l))(X_n - E(X_n))] \\ \vdots & \ddots & \vdots \\ E[(X_n - E(X_n))(X_l - E(X_l))] & \cdots & E[(X_n - E(X_n))(X_n - E(X_n))] \end{bmatrix} \tag{5}$$

Based on the covariance matrix, the eigenvalues and the corresponding eigenvectors are calculated, and then the eigenvectors are unitized and arranged in the order of the eigenvalues from the largest to the smallest. This set of mutually orthogonal unit vectors is the so-called principal components, and in order of the order of the first principal component, the second principal component and so on. Moreover, since the covariance matrix is a real symmetric matrix, the eigenvectors corresponding to different eigenvalues are orthogonal to each other, so it is a set of orthogonal unit vectors, which is also called standard orthogonal basis.

This decomposition method is also known as eigenvalue decomposition, which can be defined as:

$$D = Q \Lambda Q^T = Q \begin{bmatrix} \lambda_l & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} Q^T \tag{6}$$

where D is the real symmetric matrix of covariance, Q the standard orthogonal matrix consisting of eigenvectors, and λ_i the diagonal matrix consisting of eigenvalues. In principal component analysis, the desired vector can also be obtained based on singular value decomposition, which is defined as:

$$A = U \Sigma V^T = U \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} V^T \tag{7}$$

where A is the $M \times N$-order matrix, U is the $M \times M$-order You matrix, V is the $N \times N$-order unit orthogonal array, and $\sum$ is the singular value matrix.

After obtaining the principal components, the next step is to obtain the principal component score, denoted as S. The principal components can be viewed as a set of basis vectors, denoted as C . And the original all centered data D can be obtained from the linear variation of all basis vectors, so it can be written as:

$$D = S \times C^T \tag{8}$$

where the principal component score is actually the weight of each principal component in making up each set of original spectra, or it can be said to be the proportion of each principal component in each line of spectral data. According to Eq:

$$S = D \times (C^T)^{-1} \tag{9}$$

And since the inverse of an orthogonal matrix is equal to its transpose, it is said:

$$S = D \times C \tag{10}$$

II. B.Fama-Macbeth regression

The study looks at the ability of the factors to explain stock returns in the cross-section from a regression perspective and after the portfolio ranking method this study introduces the Fama-Macbeth regression. In testing the asset pricing process, factor returns and factor exposures are examined through a two-step regression from factor exposures.

Fama-Macbeth regression [28] is performed in two steps. In the first step let the individual stock returns be regressed on the factor values in a time series to obtain the factor exposure. In the second step the factor exposure obtained in the previous step is used as the dependent variable and regressed on each cross-section, the regression results for each period are averaged over the time series and tested to see if the mean is significantly different from zero, which is used to confirm whether the factor has a systematic impact on the return on assets over the entire time period. The mathematical expressions are as follows:

$$R_t^{ei} = a_i + \beta_i^{'} f_t + \grave{o}_t^i, \quad t = 1, 2, \ldots, T \tag{11}$$

$$R_t^{ei} = \hat{\beta}_i^{'} \lambda_t + \alpha_{it}, \quad i = 1, 2, \ldots, N \tag{12}$$

III. Level, slope and curvature factor pricing modeling

This section uses the method of extracting pricing factors from stock returns to construct a level, slope, and curvature factor model. The first step is to estimate the expected returns of individual stocks using firm characteristic variables as predictor variables and to form multiple portfolios using the expected returns sorting. The second step is to extract the pricing factors from the portfolio returns, which are level, slope, and curvature factors, through the principal component analysis method, and construct the level, slope, and curvature factor pricing model (LSC model). This method extracts the common fluctuations that are related to the expected returns of the assets, while removing the perturbations that are not related to the expected returns.

III. A. Single Dimension Portfolio Construction Methodology

In this paper, we use a three-step approach to realize the construction of portfolios using expected returns as a single dividing metric, utilizing multiple firm characteristics as predictor variables, deriving the expected return of each stock through Fama-Macbeth cross-sectional regression, and then ranking the firms' stocks according to the magnitude of their expected returns, which leads to the formation of portfolios.

The paper proceeds as follows: first, the predictor variable of the i th stock at time t is used as the independent variable, and the portion of the return of the i th company's stock that exceeds the risk-free rate of interest, i.e., the excess return of the i th company's stock (XRet), is used as the dependent variable, and the parameters of the model are estimated using Fama-Macbeth cross-sectional regression.

The specific form of the cross-sectional regression required for stocks at each time t is as follows:

$$\begin{aligned} XRet_{i,t+1} = {} & \beta_0 + \beta_1 L\log Size_{i,t} + \beta_2 L\log B/M_{i,t} + \beta_3 Mom_{i,t} + \beta_4 zeroNS_{i,t} \\ & + \beta_3 NS_{i,t} + \beta_6 NegACC_{i,t} + \beta_7 posACC_{i,t} + \beta_8 dA/A_{i,t} \\ & + \beta_9 posOP_{i,t} + \beta_{10} negOP_{i,t} + \varepsilon_{i,t+1} \end{aligned} \tag{13}$$

where $XRet_{i,t+1}$ is the excess return of the i th stock at the $t+1$ th moment. $LogSize_{i,t}$ is the natural logarithm of the stock's outstanding market capitalization. $LogB/M_{i,t}$ is the natural logarithm of the book-to-market ratio of the stock. $Mom_{i,t}$ is the momentum factor, which is the cumulative compounded return on stock i from month $t-12$ to month $t-2$. $NS_{i,t}$ is the stock's net issuance, divided into zeroNS and NS variables depending on whether this value is zero. $ACC_{i,j}$ is the accruals for this stock, divided into two variables, positive accrual posACC and negative accrual NegACC, based on whether the value is positive or negative. $OP_{i,t}$ is the stock's operating profitability, which is divided into positive operating profitability negOP and negative operating profitability $posOP$ based on whether the value is positive or negative. $dA/A_{i,t}$ is the asset growth rate of the stock. $\varepsilon_{i,\mu+1}$ is the random perturbation term of the regression.

The study utilizes the estimated coefficients derived from cross-sectional regressions [29] performed on stocks to predict the return of each stock for the coming month, i.e., to obtain a unique indicator of the expected return that divides the portfolio. Where the expected return obtained for each stock is the fitted value of its regression. Since the sample stocks are divided into three different size groups based on size, the estimated coefficients obtained from the cross-sectional regression for each size group may differ somewhat. Returns are forecasted for each stock in the different size groups. Using the predicted returns as the grouping criterion for the portfolios, a one-dimensional grouping method is used to rank the expected returns of the selected stocks from smallest to largest, and the stocks are divided into 25 portfolios on average based on the 0.04, 0.08, ..., 0.92, and 0.96 quartile points of the expected returns. In this case, the portfolio division is done once a month at the end of the month, meaning that 25 portfolios are constructed and formed every month.

III. B. Extraction of pricing factors

Principal component analysis utilizes eigenvalue decomposition to extract common influencing factors, i.e., principal component factors, from multiple variables, and effectively reflects the information of the original variables to better capture the intrinsic patterns among variables. In this paper, this method is used to extract the public pricing factors that affect asset returns from portfolio returns, which are uncorrelated and each factor is a linear combination of portfolio returns. In addition, since the portfolio is constructed by using the expected return as a dividing indicator, the perturbations that are not related to the expected return are removed, and the principal component analysis of its return can extract the public pricing factors that are only related to the expected return, and then the level, slope, and curvature factor model is constructed. In addition, in many empirical analyses, principal component analysis can be used to eigenvalue decompose the correlation matrix, or by standardizing the variables and then eigenvalue decomposing the covariance matrix. In this case, the variables are standardized in order to be able to eliminate possible effects between the variables due to inconsistencies in the scales.

III. C. Comparative Pricing Models

In this paper, we compare the pricing models choosing the well-known CAPM model, FF-3 factor model, Carhart-4 factor model, and CH-3 factor model.

(1) The CAPM model decomposes the return on assets into two components: the market risk-free return and the risk premium gained from taking risks, which is given by the formula:

$$E(r_i) = r_f + \beta_{im}(E(r_m) - r_f) \tag{14}$$

(2) Fama-French three-factor asset pricing model [30], based on the CAPM, add the market capitalization factor SMB, book-to-market ratio factor HML to form a three-factor model (FF-3). The expression is as follows:

$$R_{it} - R_{ft} = \alpha + \beta_{im} RMRF_t + \beta_{is} SMB_t + \beta_{ih} HML_t + \grave{o}_t \tag{15}$$

(3) Carhart-4 Factor Pricing Model, which adds a momentum factor to the three-factor model, taking into account the impact of systematic risk, book-to-market ratio, market capitalization size and momentum factors on stock performance. The current Carhart four-factor model (Carhart-4) was developed to comprehensively evaluate fund performance and historically effectively measure a fund's ability to achieve excess returns through active investment management.

(4) CH-3 factor pricing model, considering that the existing pricing models are based on the asset returns of foreign markets. While this study tests the factors for the A-share market, the pricing model controlled is needed to explain more about the Chinese A-share market. After considering the IPO institutional arrangement in the Chinese market, small stocks with shell value contamination are excluded, the ranking indexes of the value factor are determined, and the FF-3 factor model is altered. EP (price-earnings ratio) is used instead of BM (book-to-market ratio) as the ranking indicator of the value factor, and a factor model that can be used for pricing in the Chinese A-share market is constructed.

III. D. Data sources and processing

III. D. 1) Data sources

In this paper, the monthly frequency data of China's A-share market is selected. Specifically, the monthly frequency data of selected sector-specific constituents are based on the industry classification of the Securities and Futures Commission (SFC), considering that the excess returns of individual stocks may be correlated with the industries in which they are located. The Chinese stock market underwent profound changes in 2008 due to the global financial crisis, and the inconsistency of the pre- and post-crisis regimes may pose an unknown risk to the model. Therefore, the pharmaceutical manufacturing industry is used as an example to focus on the asset pricing of specific industries in China in the post-crisis era. The start time of the stock data is chosen to be October 2008 after the end of the financial crisis as the initial point, and the end point is chosen to be January 2025. In order to be consistent with the real investment decision-making process, retain the time-series features in the data, and utilize the sliding window method for rolling forecasts, this paper selects 117 company characteristic variables to proxy for the anomaly factors, including 42 volume and price indexes and 75 financial fundamentals of the company, and the data are obtained from the Guotai Junan database, CSMAR.

III. D. 2) Data processing

Data processing, consistent with the existing literature, is divided into three parts for data processing, extreme value processing, missing value processing and data standardization.

(1) Extreme value processing

Financial data is always limited by extreme values, the existence of extreme values may have an excessive effect on the results, making the results not credible. Shrinking tail processing is to set the value greater than the upper h percentile of all observations to the 100-h percentile, and the value less than the k percentile under the observation to the k percentile, this paper adopts shrinking tail processing of extreme values, h, k are taken as 0.5.

(2) Missing value processing

For quarterly financial data are filled monthly: in general, listed companies publish their annual reports after May, semi-annual reports after September, and quarterly statements are often published at different times. For the missing financial data of the current month, the latest financial data published in the previous period will be populated.

For missing return data: if a stock has missing return data in month t, exclude all data on month t for that stock for missing indicators of volume and price: if a stock's factor value is missing, fill it with 0.

(3) Normalization

Due to inconsistent data magnitudes in the set of anomalous factors, the variance of one factor may be a higher-order quantity of the variance of other factors, and thus dominate the model predictions, which may not be due to the fact that the factor itself can significantly affect the expected returns, but simply due to the data magnitude, and may cause the model to ignore the really important characteristic factors. Therefore, data standardization can make features comparable across dimensions and magnitudes, and can improve computational efficiency and convergence speed. For the above considerations, the data in the set of anomaly factors are standardized.

Because the sample selected in this paper belongs to the pharmaceutical manufacturing industry, the steps of excluding stocks in the financial industry and neutralizing the factors by industry are omitted. After the above data processing, the final valid samples are 41,359.

IV. Empirical testing

IV. A. Fama-Macbeth regression test

The relationship between credit risk (PD) and portfolio excess returns is first examined without introducing control variables, and then control variables characterizing the level of market capitalization and valuation of the stock (sizeSize to book-to-market ratio BM), an indicator characterizing the level of liquidity of the stock (turnover rate TURN), and other control variables characterizing the information on the historical returns of the stock are added gradually. Considering the magnitude of the variables, the natural logarithm is taken for the size of the individual stock (SIZE) and the distance to default is used to measure credit risk. Table 1 shows the results of the Fama-Macbeth regression.

From the results of model 1 in the regression results, it can be seen that the coefficient of the PD term is -0.473, which is significantly negative without introducing other control variables. It indicates that there is a significant negative correlation between credit risk and stock excess return, i.e., there is a negative premium for credit risk.

Meanwhile, from Model 2 to Model 4, it can be seen that after gradually adding the control variables such as size level, book-to-market ratio, turnover rate and momentum indicator, the coefficients of the PD term are always negative and all of them are significant at 99% confidence level, indicating that the negative correlation between credit risk and stock excess return is very robust.

It indicates that investors tend to include the assessment of credit risk of listed companies when analyzing the value of assets, and investors are more reluctant to invest in high credit risk companies. This is because such firms have a higher likelihood of defaults, and in addition, since the law requires that debt investors have a higher compensation priority than equity investors, which can cause greater losses to equity investors, investors would prefer to buy shares of underlying companies with good performance and lower risk rather than taking on a portion of the credit risk in order to obtain a risk premium.

Table 1: Results of Fama Macbeth's return

	Model 1	Model 2	Model 3	Model 4	Model 5
PD	-0.473***(0.126)	-0.402***(0.113)	-0.243***(0.085)	-0.162***(0.067)	-0.115***(0.042)
Size		-0.043***(0.013)	-0.027***(0.007)	-0.010***(0.005)	-0.023***(0.041)
BM			0.004***(0.003)	0.007(0.006)	0.011**(0.005)
TURN				0.462(1.144)	-0.009(0.002)
MOM					0.019***(0.037)
_cons	0.002***(0.007)	0.007***(0.006)	0.005***(0.004)	0.009***(0.004)	0.003***(0.002)

Note: *** indicates that the model passes the 0.001 level of significance test.

IV. B. Portfolio descriptive statistics

The Fama-Macbeth regression gives the average returns of the 25 portfolios and the corresponding projected returns. Each month's portfolio is the predicted return from a cross-sectional Fama-Macbeth regression of each firm's return over the risk-free rate on the same size, book value, momentum, stock's net issuance for the year, accrued expenses, asset growth, operating margins, and turnover rate. Because the momentum characteristics need to be updated monthly, the regressions need to be run every month. Based on the predicted returns, firms can be categorized into 25 market capitalization-weighted portfolios. In this case, the predicted returns are the average fitted values weighted by market capitalization outstanding, while the average returns are the average returns weighted by the market capitalization outstanding of the companies in each portfolio.

Table 2: Variables of each portfolio

	AEI	EEI	MOM	BM	TURN	Log(Size)	INV	ACC	OP	NS
1	-0.46	-0.29	-0.46	-0.03	0.53	10.35	0.89	0.00	-0.37	0.09
2	-0.43	-0.28	-0.43	0.02	0.76	10.51	0.89	0.00	-0.03	0.09
3	-0.37	-0.26	-0.39	0.04	0.85	10.70	0.89	0.00	0.00	0.07
4	-0.33	-0.22	-0.38	0.04	1.13	10.89	0.89	0.00	0.00	0.07
5	-0.27	-0.18	-0.35	0.07	1.40	10.95	0.90	0.00	-0.01	0.07
6	-0.25	-0.15	-0.30	0.09	1.54	11.17	0.92	0.00	0.02	0.06
7	-0.23	-0.12	-0.23	0.10	1.68	11.33	1.00	0.00	0.02	0.05
8	-0.19	-0.10	-0.14	0.11	1.72	11.49	1.00	0.00	0.02	0.04
9	-0.18	-0.08	-0.07	0.11	1.95	11.65	1.00	0.00	0.02	0.01
10	-0.16	-0.07	0.00	0.14	2.18	11.82	1.00	0.00	0.04	0.01
11	-0.14	-0.06	0.04	0.19	2.50	11.98	1.00	0.00	0.04	0.01
12	-0.10	-0.03	0.07	0.21	2.73	12.21	1.00	0.00	0.04	0.01
13	-0.05	-0.02	0.11	0.26	3.01	12.40	1.00	0.00	0.04	0.01
14	-0.03	0.01	0.15	0.35	3.19	12.54	1.01	0.01	0.04	0.00
15	0.05	0.01	0.21	0.39	3.42	12.77	1.01	0.01	0.06	0.00
16	0.07	0.02	0.27	0.42	3.60	13.19	1.02	0.01	0.06	0.00
17	0.09	0.04	0.32	0.45	3.83	13.36	1.02	0.01	0.06	0.00
18	0.10	0.06	0.35	0.47	4.11	13.50	1.02	0.02	0.06	0.00
19	0.13	0.10	0.41	0.52	4.34	13.68	1.03	0.02	0.07	0.00
20	0.18	0.13	0.48	0.65	4.52	13.96	1.03	0.05	0.07	0.00
21	0.21	0.18	0.58	0.69	4.85	14.17	1.03	0.07	0.08	-0.01
22	0.21	0.23	0.63	0.74	5.12	14.34	1.03	0.07	0.10	-0.01
23	0.24	0.27	0.72	0.83	5.35	14.48	1.03	0.14	0.11	-0.02
24	0.27	0.31	0.76	1.03	5.58	14.71	1.03	0.17	0.13	-0.03
25	0.31	0.34	0.88	1.18	10.95	15.15	1.09	0.26	0.15	-0.08

Table 2 shows the descriptive statistics of each variable under the 25 sorted portfolios. Namely, the descriptive statistics of the average excess return AEI and the expected excess return EEI obtained through the model for each portfolio, momentum MOM, book-to-market ratio BM, turnover ratio TURN, log firm size Log(Size), investment INV, accrued expenses ACC, operating profit margin OP and net stock issuance NS.

The characteristics of each portfolio are obtained by weighting the market capitalization outstanding by the return of each stock in the portfolio so that the portfolio is not dominated by a large but tiny number of small-cap stocks. Such a sorting methodology makes a large difference in average excess returns, similar to the difference in predicted excess returns. Almost all of the multiple regression metrics show a monotonic pattern of growth or decline in the same or opposite direction as expected returns. It is worth noting that there is an opposite trend between the annual net issuance NS of stocks and expected returns, i.e., low expected returns for high net issuance portfolios and higher expected returns for low net issuance portfolios, suggesting that a company's additions to its stock portfolio lead to a decline in long-term stock returns, whereas stock repurchases lead to an increase in the long-term returns of the company's stock. The Accruals ACC has a large gap between the extreme high return and extreme low return groups, while the growth rate is slower between the middle groups with large standard deviations.

Figure 1 shows a graph of the average excess returns and expected excess returns of the 25 portfolios, and it can be seen that the asset pricing method chosen in this paper has a better predictive effect, the predicted returns are much closer to the true returns, and the 25 portfolios show a smoother upward trend.

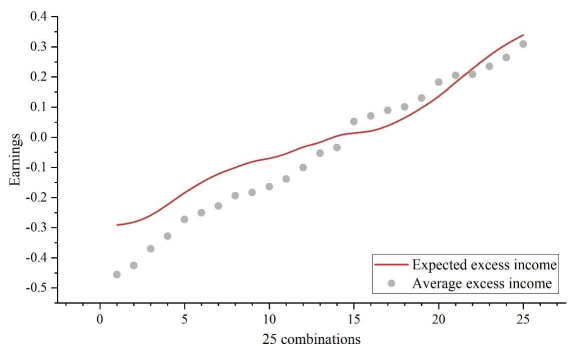

Figure 1 Average excess earnings and expected excess earnings

IV. C. Factor extraction under the PCA method

Through the regression in the previous step, we obtained the expected return, average return situation and each index worth situation of each portfolio, and the next step is to extract the public factors from these portfolios. Using the prediction results of the anomaly regression, the 25 portfolios sorted by expected return from low to high are subjected to principal component analysis, and the eigenvalues and variance explanations of each principal component obtained are shown in Table 3. The table shows the first ten principal components, including their respective eigenvalues and the variance explained. The first three components explain 95% of the variance of the portfolios, which are 90.3%, 3.7% and 1.4%, respectively.

Table 3: The analysis of the main component analysis of 25 abnormal sorting

Principal component	eigenvalue	Simple variance interpretation	Cumulative method interpretation
1	22.575	0.903	0.903
2	0.925	0.037	0.94
3	0.35	0.014	0.954
4	0.2	0.008	0.962
5	0.15	0.006	0.968
6	0.125	0.005	0.973
7	0.1	0.004	0.977
8	0.075	0.003	0.98
9	0.05	0.002	0.982
10	0.05	0.002	0.984

Figure 2 shows the factor loadings of the first three principal components of the 25 abnormally sorted portfolios. The first principal component is similar to the market portfolio, with roughly equal weights across the 25 portfolios, so it can be called the “level” factor, which explains about 90% of the variance in the 25 portfolios. The second principal component is a long position in stocks with low expected returns and a short position in stocks with high expected returns. Its loadings are essentially monotonically increasing. On average, it has a negative coefficient. Because of its shape, it is called the “slope” factor. The slope factor focuses on the expected return of the stock corresponding to the company, rather than on underlying variables of the expected return at the company level, such as the size of the company, the profitability of the company, and so on. The third principal component is to be short the extreme low expected return and high expected return portfolios while being long the intermediate portfolios. The factor loadings of this factor have a parabolic shape, so it is called the “curve” factor. This factor indicates that extreme stocks tend to move in the same direction, i.e., they tend to move in the same direction up or down.

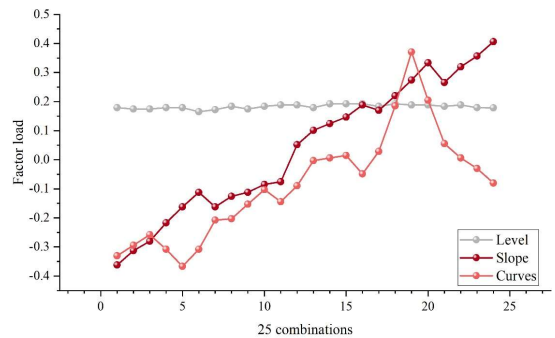

Figure 2: The factor load of the first three main components of the portfolio

IV. D. Comparison of pricing power of factor models

In this section, the explanatory power of all asset pricing factor models is empirically evaluated, including the CAPM model, the FF-3 model, the Carhart-4, the CH-3 model and the LSC model in this paper. Taking the above 25 binary grouped portfolios as samples, the regression analysis is conducted with the excess returns of the 25 portfolios as the explanatory variables and the different combinations of the factors as the explanatory variables. Meanwhile, this paper refers to the previous literature research results and uses the following four statistics as a measure of the explanatory power of the factor model.

(1) GRS test: the original hypothesis of GRS test is that the constant term in multiple regressions is jointly equal to zero, when the GRS statistic is greater than the critical value, the original hypothesis is rejected, which can be used in asset pricing models, where the portfolio return is used as the explanatory variable in asset pricing models, and each pricing factor is used as the explanatory variable. If the GRS test statistics done in groups are greater than the critical value, it means that the original hypothesis that the intercept term in multiple regressions is jointly equal to zero is rejected and the pricing model does not explain the return of the asset portfolio well, and vice versa, the original hypothesis is accepted, which means that the asset pricing model explains the return in a respectable way.

(2) $A|a_i|$: the mean of the absolute value of the constant term, when doing the regression for each portfolio, the smaller the constant term indicates the smaller the unexplained portion of the portfolio return, so the smaller this mean is, the stronger the explanatory power of the factor pricing model.

(3) $A|a_i|/A|\overline{r_i}|$: the ratio of the mean of the absolute value of the constant term to the mean of the degree of deviation of the portfolio return, where the degree of deviation of the portfolio return is defined as the difference between the portfolio return and the mean of the return. The smaller $A|a_i|/A|\overline{r_i}|$ is the smaller the unexplained portion of the return variance is, and the more impressive the explanatory power of the pricing model.

(4) $A(Adj.R^2)$: the average value of the adjusted R^2 of the regression results, in linear regression R^2 is known as the regression's goodness of fit, the larger the R^2 the better the model fits the sample data, but the increase in the number of independent variables will significantly increase the value of R^2, so this paper chose to adjust R^2 to measure the model's fitting ability. A larger $A(Adj.R^2)$ indicates a stronger factor explanatory power.

The comparison of the explanatory power of the multi-factor pricing model is shown in Table 4. From the test results of GRS, it can be found that the GRS statistics of CAPM model, FF-3 model, Carhart-4 model, and CH-3 model significantly reject the original hypothesis. That is, the intercept term of all regressions is jointly not equal to zero and the pricing effect is limited. In contrast, the GRS statistic under the LSC model in this paper is 1.243 and

insignificant, indicating that the original hypothesis is accepted and the intercept term of the regression is jointly equal to zero, which indicates that the LSC in this paper has good explanatory power for excess return.The GRS statistic of the LSC model is lower compared to the baseline model by 0.286-0.930, which indicates that the LSC model in this paper has better explanatory power.

In addition, observing the values of the other three indicators, it can be found that the $A(Adj.R^2)$ statistic of this paper's model is 0.941, which is improved by 0.162~0.270 compared with that of the baseline model, indicating that this paper's model has a better ability to fit the sample. The $A|a_i|$ and $A|a_i|/A|\overline{r_i}|$ statistic also roughly shows that the LSC model is significantly lower than the baseline model. It can be concluded that the LSC model in this paper has better explanatory power compared to other factor pricing models.

Table 4: Comparison of explanation ability of multi-factor pricing model

Models	GRS	$A\|a_i\|$	$A\|a_i\|/A\|\overline{r_i}\|$	$A(Adj.R^2)$
CAPM	2.173*	0.0093	1.317	0.671
FF-3	1.957*	0.0076	1.224	0.694
Carhart-4	1.864*	0.0059	1.075	0.735
CH-3	1.529*	0.0043	0.953	0.779
LSC(Ours)	1.243	0.0021	0.372	0.941

V. Conclusion

The study constructed an asset pricing model for corporate data based on level, slope and curvature, mined and processed multiple corporate asset stock market data. It is also divided into multiple portfolios according to expected returns, and then combined with principal component analysis to extract asset pricing factors from them. In addition, the relationship between corporate risk and expected return is examined by Fama-Macbeth regression in order to statistically analyze the average return and other indicators under each portfolio. The study concludes by comparing the asset pricing power of different models and evaluating the strength of the models' pricing explanations based on tests such as GRS.

The correlation coefficient between credit risk and excess return on issuance in the Fama-Macbeth test is -0.473, indicating that credit risk has a negative premium effect on asset pricing. As the market capitalization of the stock yield outstanding increases incrementally, the annual net issuance of stocks shows an opposite trend between the net issuance of stocks and the expected excess return. The explanation of the first ten principal components of the portfolio exceeds 98%, with the strongest component explaining 90.3% of the total.The GRS statistic of the LSC model is insignificant, the $A(Adj.R^2)$ statistic improves by 0.162 to 0.270 from the baseline model, and the $A|a_i|$ and $A|a_i|/A|\overline{r_i}|$ statistics are reduced to different degrees from the baseline model, i.e., the LSC model has a better ability to explain excess returns.

References

[1] Jiang, H., & Murmann, J. P. (2022). The rise of China's digital economy: An overview. Management and Organization Review, 18(4), 790-802.

[2] Xu, Y., & Li, T. (2022). Measuring digital economy in China. National Accounting Review, 4(3), 251-272.

[3] Qi, W., Sun, M., & Hosseini, S. R. A. (2023). Facilitating big-data management in modern business and organizations using cloud computing: a comprehensive study. Journal of Management & Organization, 29(4), 697-723.

[4] Pei, J., Zhu, F., Cong, Z., Luo, X., Liu, H., & Mu, X. (2021, August). Data pricing and data asset governance in the AI era. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (pp. 4058-4059).

[5] Hannila, H., Silvola, R., Harkonen, J., & Haapasalo, H. (2022). Data-driven begins with DATA; potential of data assets. Journal of Computer Information Systems, 62(1), 29-38.

[6] Li, Y. (2024). Concepts, Accounting Treatment and Pricing of Data Assets. Journal of Economic Insights, 1(1), 7069393-7069393.

[7] Tang, Q., Shao, Z., Huang, L., Yin, W., & Dou, Y. (2020). Identifying influencing factors for data transactions: a case study from shanghai data exchange. Journal of Systems Science and Systems Engineering, 29, 697-708.

[8] Goldfarb, A., & Tucker, C. (2019). Digital economics. Journal of economic literature, 57(1), 3-43.

[9] Pei, J. (2020). A survey on data pricing: from economics to data science. IEEE Transactions on knowledge and Data Engineering, 34(10), 4586-4608.

[10] Lin, J., Huang, Z., & Tang, Y. (2025). Pricing for Data Assets Based on Data Quality, Quantity and Utility on the Perspective of Consumer Heterogeneity. IEEE Transactions on Knowledge and Data Engineering.

[11] Heckman, J. R., Boehmer, E. L., Peters, E. H., Davaloo, M., & Kurup, N. G. (2015). A pricing model for data markets. IConference 2015 Proceedings.

[12] Yu, H., & Zhang, M. (2017). Data pricing strategy based on data quality. Computers & Industrial Engineering, 112, 1-10.

[13] Carter, T. (2007). An introduction to information theory and entropy. Complex systems summer school, Santa Fe, 15-34.

[14] Rodríguez, R. A., & Cáceres-Hernández, J. J. (2018). Information, entropy, value, and price formation: An econophysical perspective. Physica A: Statistical Mechanics and its Applications, 512, 74-85.
[15] Shen, Y., Guo, B., Shen, Y., Duan, X., Dong, X., & Zhang, H. (2019, January). Pricing personal data based on information entropy. In Proceedings of the 2nd International Conference on Software Engineering and Information Management (pp. 143-146).
[16] Li, X., Yao, J., Liu, X., & Guan, H. (2017, June). A first look at information entropy-based data pricing. In 2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS) (pp. 2053-2060). IEEE.
[17] Dong, J. I. A. N. G., Xiaowei, Z. H. A. N. G., & Ye, Y. U. A. N. (2023). A model game pricing method based on data purity. Frontiers of Data and Domputing, 5(5), 63-73.
[18] Bin, Z., Jiqun, Z., Wei, C., Bin, L., Wenyin, Z., & Yilong, G. (2023). A dynamic evaluation model of data price based on game theory. Peer-to-Peer Networking and Applications, 16(5), 2073-2088.
[19] Koutris, P., Upadhyaya, P., Balazinska, M., Howe, B., & Suciu, D. (2013, June). Toward practical query pricing with querymarket. In proceedings of the 2013 ACM SIGMOD international conference on management of data (pp. 613-624).
[20] Xiao, Z., He, D., & Du, J. (2020). A Stackelberg game pricing through balancing trilateral profits in big data market. IEEE Internet of Things Journal, 8(16), 12658-12668.
[21] Chen, Y., Liu, M., Zhang, J., Tan, A., & Wang, Y. (2024). A Three-Party Dynamic Pricing Mechanism for Customized Data Products Based on the Stackelberg Game and Bargaining Model. Symmetry, 16(11), 1540.
[22] Hu-Bolz, J., Reed, M., Zhang, K., Liu, Z., & Hu, J. (2025). Federated data acquisition market: Architecture and a mean-field based data pricing strategy. High-Confidence Computing, 5(1), 100232.
[23] Bi, Y., Wu, Y., Liu, J., Ren, K., & Xiong, L. (2024, May). When Data Pricing Meets Non-Cooperative Game Theory. In 2024 IEEE 40th International Conference on Data Engineering (ICDE) (pp. 5548-5559). IEEE.
[24] Li, J., Li, J., Wang, X., Qin, R., Yuan, Y., & Wang, F. Y. (2023). Multi-blockchain based data trading markets with novel pricing mechanisms. IEEE/CAA Journal of Automatica Sinica, 10(12), 2222-2232.
[25] Oh, H., Park, S., Lee, G. M., Choi, J. K., & Noh, S. (2020). Competitive data trading model with privacy valuation for multiple stakeholders in IoT data markets. IEEE Internet of Things Journal, 7(4), 3623-3639.
[26] Fernandez, R. C., Subramaniam, P., & Franklin, M. J. (2020). Data market platforms: Trading data assets to solve data problems. arXiv preprint arXiv:2002.01047.
[27] Lingbin Shen,Liping Tian,Hongbing Yao,Dongpeng Tian,Yifan Ge,Zhongmou Sun & Yuzhu Liu. (2024). Rapid detection and recognition of phosphors using laser-induced breakdown spectroscopy and principal component analysis method—back propagation neural network algorithm. Laser Physics,34(5),
[28] Kyuseok Lee. (2020). A weighted Fama-Macbeth two-step panel regression procedure: asymptotic properties, finite-sample adjustment, and performance. Studies in Economics and Finance,ahead-of-print(ahead-of-print),347-360.
[29] Sikai Xu,Jianping Hu,Zhiyi Ouyang,Maolin Yuan,Yan Zheng,Xin Liu & Yang Shen. (2025). Elevated atherogenic index of plasma is associated with increased cardiorenal syndrome prevalence: a cross-sectional study. Renal failure,47(1),2472037.
[30] Mohammad Q.M. Momani. (2021). On the robustness of the Fama-French three-factor and the Carhart four-factor models on the Amman Stock Exchange. Afro-Asian Journal of Finance and Accounting,11(1),
Mark J. Kamstra & Ruoyao Shi. (2024). Testing and Ranking of Asset Pricing Models Using the GRS Statistic. Journal of Risk and Financial Management,17(4),168-.