International Review of Economics and Finance 101 (2025) 104153

Contents lists available at ScienceDirect

International Review of Economics and Finance

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

The pricing ability of factor model based on machine learning: Evidence from high-frequency data in China

Ailian Zhang [a], Mengmeng Pan [b], Xuan Zhang [c,*]

[a] School of Business and Management, Jilin University, No.2699 Qianjin Street, Changchun 130012, China
[b] School of Management and Economics, The Chinese University of Hong Kong, Shenzhen, No. 2001, Longxiang Avenue, Longgang District, Shenzhen 518172, China
[c] College of Economics and Management, Nanjing University of Aeronautics and Astronautics, Nanjing, China

ARTICLE INFO

JEL classification:
E52
G28
E31
E32

Keywords:
Asset pricing
Machine learning
High-frequency
Chinese stock market

ABSTRACT

The existing literature mainly documents the asset pricing models estimated on low-frequency data, lacking the empirical evidence for exploring the "right" systematic factors based on high-frequency (HF) level. This study develops a revised HF factor model and evaluates the asset pricing performance. Using machine learning algorithms, we find that HF factor model includes three very persistent systematic factors, well-approximated by a portfolio of market, finance, and information. Sharpe ratios and out-of-sample tests prove that the HF revised factor model has the best explanatory power compared to the CAPM, Fama-French three-factor and five-factor models. The findings contribute to an in-depth understanding of the characteristics and mechanisms of risk and return from an HF perspective in the Chinese stock market.

1. Introduction

As a crucial component of contemporary financial market pricing theory, asset pricing models focus on revealing the relationship between asset returns and risk, and finding the "right" systematic factors helps to improve the explanatory power of the model (Cayirli et al., 2022; Pelger, 2020). Traditional research mainly relies on low-frequency data to construct asset pricing models (Hendershott et al., 2022; Herskovic et al., 2023), thereby overlooking the issue of information gaps inherent in such data and subsequently diminishing the explanatory power of the models (Mallinger-Dogan & Szigety, 2014). Specifically, low-frequency data possess characteristics of opacity and lag, rendering the constructed models unable to explain intraday risks (Andersen et al., 2023; Louzis et al., 2013), and fail to account for the heterogeneity patterns in stock price fluctuations during the day from opening to closing. Therefore, identifying and measuring systematic factors from a HF perspective facilitates the construction of accurate pricing models.

Nowadays, numerous empirical and theoretical studies have proposed the systematic role of HF data in explaining asset risk and return sensitivity (Lettau & Pelger, 2020; Renault, 2017; Wang, Chen, Lian, & Chen, 2022; Zhu et al., 2017). However, related researches mainly focus on developed countries (Cartea & Jaimungal, 2013; Hollstein et al., 2020; Pelger, 2020). This paper mainly explores the applicability of HF models in the stock market of the largest developing country - China. The main reasons are as follows: Firstly, there are differences in the composition and performance of risk factors under different securities market systems. Developed countries' stock markets follow the "T+0" trading rule, while China's stock market follows the "T+1" trading rule, which means that

* Corresponding author.
E-mail addresses: ailianzh@jlu.edu.cn (A. Zhang), panmengmeng@cuhk.edu.cn (M. Pan), xuanzhang.nanking@gmail.com (X. Zhang).

https://doi.org/10.1016/j.iref.2025.104153
Received 4 August 2024; Received in revised form 18 March 2025; Accepted 5 May 2025
Available online 6 May 2025

stocks bought on the first trading day can only be sold after the second trading day. The trading intervals help investors fully absorb market information, and meanwhile facilitate government policy intervention in the market, leading to uncertainty in stock market volatility and requiring more accurate analysis and judgment (Brunnermeier et al., 2022; Qiao & Dam, 2020). Models constructed from HF data can better address this challenge. Secondly, according to data analyzed by the Securities Industry and Financial Markets Association (SIFMA), China has emerged as the world's third-largest stock market since 2023, with a market share of 10.6 % of the global stock markets. Considering the growth and significance of the Chinese stock market, it is imperative to explore the construction of an asset pricing model for the Chinese stock market.

Existing studies on stock markets of emerging country are limited by the difficulty of accessing and processing of HF data, and rarely simultaneously consider both HF data and all individual stocks in the stock market (Yao et al., 2022; Zhang et al., 2021). With the advancements in big data technology, the utilization of machine learning techniques become feasible for quantifying the intraday effects of stock markets and enhance the explanatory power of pricing models (Leippold et al., 2022). Specifically, machine learning techniques can effectively handle large-dimensional data, enable automated discovery of latent patterns hidden in the data, and effectively address intricate nonlinear problems. Therefore, this study employs machine learning principal component analysis (PCA) method to extract valuable information from the HF 5-min data of all A-shares in the stock market, aiming to construct a HF statistical factor model.

Statistical factors are difficult to analyze the intrinsic economic logic. The main reason is that existing research is based on factor sets with economic significance to extract principal components and construct statistical factors, such as size, market value, or ESG investment variables (Bergbrant & Kelly, 2016; Guobužaitė & Teresienė, 2021). So, it is difficult to explain the complex economic implications of statistical factors. Unlike existing research, this paper extracts value information from all individual stock volatility data to form statistical factors, and its source is not from existing economic variables. Afterwards, the GC method is used to search for potential factors that are similar to the volatility of the statistical factor, which can approximately indicate the source of the actual factors that cause stock volatility. The HF model is not only beneficial for the interpretation and application of statistical factor models, but also helps to explore the practical influencing factors of stock market returns.

This paper contributes to existing literature in the following ways. First, innovation of methods. Existing studies utilize machine learning, such as neural network (Chen et al., 2023), random forest (Ma et al., 2021) and Markov classes (Grillini et al., 2019), to forecast stock market returns. This paper adopts machine learning to explore the "right" systematic factors and construct a new pricing model, aiming to improve the explanatory power of the model. Second, innovation of data. Previous literature analyzing HF asset pricing models predominantly focuses on estimating effective factors based on observable factor sets (Pástor et al., 2022), such as the three-factor model (Fama et al., 1993), profitability factor (Novy-Marx, 2013), and IPO reform factor (Li & Rao, 2022). This paper overcomes the limitation of insufficient explanatory power of observable factors under a limited set of factors by extracting value information from all stocks in a large cross-section and synthesizing comprehensive latent factors, thereby significantly enhances the explanatory power of the factors. Third, innovation in application. Extant literature mainly constructs factors by exploring the influencing factors of stock price (Bekaert & Hoerova, 2014; Fama & French, 2020), which makes each factor economically significant but prone to factors omission. This paper constructs statistical factors and then uses the GC method to find realistic proxies for the statistical factors, effectively addressing the issue of inadequate explanatory power in models due to missing factors.

The remainder of this paper is organized as follows: Section 2 is literature review, Section 3 presents the theoretical model and data processing, Section 4 presents the empirical results, Section 5 presents the results of HF risk factor model pricing capability test, Section 6 presents the results of out-of-sample test, and Section 7 presents the final section concludes.

2. Literature review

Existing research identifies and measures asset pricing factors mainly based on macro level (Guobužaitė & Teresienė, 2021; Sharpe, 1964), firm (Pástor et al., 2022), and technology levels (Kelly et al., 2021). The technical-level factor research is mainly divided into two categories. The first category is based on the set of observable factors, i.e., extracting the valid information from the existing macro or firm-level factors to fit the statistical factors (Barinov, 2014; Bergbrant & Kelly, 2016; Guobužaitė & Teresienė, 2021). However, some researchers argue that relying solely on linear regression and observable factor sets to portray the relationship between returns and risk is insufficient (Beber et al., 2015; Jenter & Lewellen, 2015; Kozak et al., 2020). Therefore, the second category involves leveraging machine learning, such as principal component analysis and incremental principal component analysis (Kelly et al., 2021), to extract and visualize value information from individual stocks in the stock market, thereby surpassing the number limitations of observable factor sets (Kozak et al., 2020; Kritzman et al., 2011).

In Chinese stock market research, the majority of scholars concentrate on the first category of research (Ma et al., 2023), while few examine the efficacy of statistical factors derived from latent factor fitting. And, given the limitations in data collection and quantification for HF statistical factors at the individual stock level, the scholars focus only on extracting latent information at the index level, using HF indices to represent all individual stocks and explain intraday effects in the stock market (Zhang et al., 2021; Yao et al., 2022). This situation results in information omission, thereby diminishing the explanatory power of the model. Therefore, this research addresses the existing lacuna.

In statistical factor models, the factors, the number of factors, and the factor loading are all unknown parameters that need to be estimated. Referring to Chamberlain and Rothschild (1983), this paper uses approximate factor models instead of traditional latent factor models to estimate unknown parameters (Pelger, 2019). The advantage lies in the fact that the approximation factor model relaxes the assumption that the trait perturbation term cannot have cross-sectional correlation, which can more effectively estimate the model parameters. Therefore, under the approximate factor model, this paper assumes that assets are independent from each other and

that the heterogeneous risk of asset returns is continuous and correlated in the cross-section.

3. Methodology

3.1. Factor estimation

The approximate factor model can be written as Eq. (1)[1]:

$$R = \Delta\widehat{F}\widehat{\Lambda}^{T} \tag{1}$$

Where R is the log of HF return in dimension M × N, $\Delta\widehat{F}$ is the estimated value of risk factors in dimension M × K, $\widehat{\Lambda}^{T}$ is the estimated value of factor loading matrix in dimension K × N. R is known, and we need to estimate $\Delta\widehat{F}$ and $\widehat{\Lambda}^{T}$. The approximate factor model has the following assumptions, $\frac{\widehat{\Lambda}^{T}\widehat{\Lambda}}{N} = I_K$, $\Delta\widehat{F}^{T}\Delta\widehat{F}$ is a diagonal matrix. After transforming Eq. (1), Eq. (2) is obtained:

$$(1/N)R^{T}R = \frac{1}{N}\widehat{\Lambda}\Delta\widehat{F}^{T}\Delta\widehat{F}\widehat{\Lambda}^{T} = \left(\widehat{\Lambda}\Big/\sqrt{N}\right)\Delta\widehat{F}^{T}\Delta\widehat{F}\setminus\left(\widehat{\Lambda}^{T}\Big/\sqrt{N}\right) \tag{2}$$

Then, $\left(\widehat{\Lambda}^{T}/\sqrt{N}\right)\left(\widehat{\Lambda}/\sqrt{N}\right) = \widehat{\Lambda}^{T}\widehat{\Lambda}/N = I_K$. Therefore, according to the definition of eigenvalue, $\widehat{\Lambda}/\sqrt{N}$ is the eigenvariable of matrix $(1/N)R^{T}R$, and the load $\widehat{\Lambda}$ is the eigenvector of $(1/N)R^{T}R^{*}\sqrt{N} = (1/\sqrt{N})R^{T}R$. And $\Delta\widehat{F}^{T}\Delta\widehat{F}$ is the diagonal matrix composed of eigenvalues. According to the definition of the unsupervised dimensionality reduction method, the larger the eigenvalue, the larger its corresponding variance, representing more information contained, so about the load is equal to taking the eigenvector corresponding to the K largest eigenvalues. Then the statistical factor is:

$$\underbrace{\Delta\widehat{F}}_{M\times\widehat{K}} = \underbrace{R}_{M\times N}\underbrace{\widehat{\Lambda}}_{N\times\widehat{K}}\left(\widehat{\Lambda}^{T}\widehat{\Lambda}\right)^{-1} \tag{3}$$

It is important to note that $\frac{1}{N}R^{T}R$ is the quadratic covariance of asset prices conditional on finite assets N. And in Eqs. (1)–(3), the observations of M and the number of assets N obey $M, N\to\infty$. This condition indicates that the values in the time series are close to continuous and the data are large samples in the cross-section, i.e., high-dimensional data, when the estimated $\Delta\widehat{F}$ and $\widehat{\Lambda}$ are consistent estimates of ΔF and Λ.

3.2. Constants and time-varying betas estimation

In estimating the time-varying beta, N-dimensional R can be represented by the following local statistical factor model:

$$R = \beta_{i,s}^{T}\Delta\widehat{F} + \Delta e, i = 1,\ldots,N, t\in(T_s, T_{s+1}), s = 1,\ldots,S \tag{4}$$

where R is the M × N-dimensional stochastic wandering process representing the logarithmic return of asset N at moment M. $\Delta\widehat{F}$ is the $M\times\widehat{K}$-dimensional stochastic wandering process known as factor return. $\beta_{i,s}$ is the $N\times\widehat{K}$-dimensional vector representing the factor coefficients and $\beta_{i,s}$ represents the exposure to the systematic factor $\Delta\widehat{F}$. The residual e_i is the K × N-dimensional random wandering process and represents the idiosyncratic risk component in addition to the systematic risk. For a time interval of $[T_s, T_{s+1}]$, β_s is a constant. This paper uses 1 month as the time window for rolling regression to calculate time-varying betas.

The constant beta for the whole sample period can be estimated as:

$$\underbrace{\beta}_{N\times\widehat{K}} = \underbrace{R^{T}}_{N\times M}\underbrace{\Delta\widehat{F}}_{M\times\widehat{K}}\left(\Delta\widehat{F}^{T}\Delta\widehat{F}\right)^{-1} \tag{5}$$

Where each asset has $\widehat{K}$ different constant beta, i.e., each systematic factor corresponds to a constant beta. Unlike the existing observable factor models, the statistical factors are based on the $\widehat{K}$ systematic risk factors obtained from PCA, and the most significant feature is that $\widehat{K}$ risk factors are orthogonal and uncorrelated with each other. Therefore, Eq. (5) can be used to calculate different constant betas and obtain consistent estimates.

3.3. Number of factors

After estimating the risk factors F, factor loadings Λ and beta coefficients, it is necessary to determine the optimal number of factors $\widehat{K}$. The selection of factor quantity is closely related to the number of eigenvalues. Referring to Ahn and Horenstein (2013) and Pelger

[1] Specific derivation steps refer to the appendix.

(2019), this paper uses perturbation method to calculate the eigenvalue ratio to determine the number of eigenvalues. Then the eigenvalues of R^TR can be found according to Eq. (1), and by arranging all the eigenvalues in order from the largest to the smallest, $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_N$ can be obtained. The perturbation method is used to find the perturbed eigenvalues as follows:

$$\widehat{\lambda}_k = \lambda_k + g(N,M), (k = 1, 2, \ldots, N) \tag{6}$$

Where $g(N,M) \in R^+$, with reference to Pelger (2019), the simulation of $g(N,M)$ value is:

$$g(N,M) = \sqrt{N} * median\{\lambda_1, \lambda_2, \ldots, \lambda_N\} \tag{7}$$

Therefore, the perturbation eigenvalue ratio statistic can be calculated as:

$$ER_k = \widehat{\lambda}_k / \widehat{\lambda}_{k+1}, (k = 1, 2, \ldots, N-1) \tag{8}$$

At this point, the perturbation eigenvalues are arranged in descending order as $\widehat{\lambda}_1 \geqslant \widehat{\lambda}_2 \geqslant \ldots \geqslant \widehat{\lambda}_N$ to calculate ER_k, and the following result is obtained:

$$\widehat{K}(\gamma) = max\{k \leqslant N-1 : ER_k \succ 1+\gamma\}, \gamma \succ 0 \tag{9}$$

Eq. (9) is the maximum value of k that is selected to satisfy the condition of $ER_k \succ 1+\gamma$.

4. Empirical analysis

4.1. Data

This paper uses 5-min stock price data for all A-shares in Chinese stock market as the initial sample, and the sample period spans from 9:30 on January 1, 2010, to 15:00 on December 31, 2020. The trading time periods of Chinese stock market are from 9:30 to11:30 and from 13:00 to 15:00 within a given trading day. For the 5-min data, there are 48 H F stock return observations per trading day, amounting to a total of 250 trading days in a year. To construct the HF risk factor, an unbalanced panel approach is employed. This approach is adopted primarily due to the inconsistent number of listed companies throughout the years covered by the sample period. Furthermore, the number of listed companies available for each year, after excluding missing values, is limited, making it impractical to generate a balanced panel dataset.

Referring to the existing literature (Cochrane, 2011; Pelger, 2020), this paper excludes the following data: (1) excludes all ST, *ST and PT stocks; (2) excludes stocks with temporary suspensions during trading days; (3) excludes stocks with less than 15 trading days in a month, and keeps the data for the first 15 days of each month for the remaining stocks. The main reason is that rolling regressions are set based on monthly 5-min HF data to calculate time-varying betas. However, holidays make it impossible to meet the consistent number of trading days per month after stock market closes, and most of the stock trading days in a year are concentrated in 18–21 days per month, but the shortest month has only 15 trading days; (4) excludes stocks with a missing rate of trading price data reaching 2.5 % within one year. Stocks with missing proportion less than 2.5 % are supplemented with the missing return data as 0. Since 0 does not provide any information when eigenvalues are extracted, it has no effect on the eigenvalue results. Stock data are obtained from the RESSET High Frequency Database.

The initial sample stock price data are log-differentiated to obtain stock return data, and the excess returns are the log return minus the risk-free rate. Referring to the existing literature (Qiao & Dam, 2020), this paper uses the Shanghai Interbank Offered Rate (SHIBOR) to measure the daily risk-free rate. The daily risk-free rate data are decomposed into 5-min data according to Pelger (2020), and the stock excess return series are the difference between 5-min HF stock return series and 5-min risk-free rate data. There are 12, 484 sample stocks in the sample period, with 8640 data for each stock in each year. Due to the unbalanced panel, the number of stocks varies each year, and the specific distribution of sample stocks in each year is shown in Table 1.

Fig. 1 shows the stationarity test results of the unit root. It shows that, from 2010 to 2020, the HF return data of all stocks remained stable at a significance level of 0.05, which means that the random components in HF data have been removed, and the changes in the data are driven by more stable factors rather than random fluctuations caused by non-stationarity, thereby removing the influence of some noise and making the data more predictable and interpretable.

4.2. Number of statistical factors

This paper mainly uses the perturbation eigenvalue ratio to select the optimal number of statistical factors. Referring to the clustering parameter method of Onatski (2010), the threshold depends on where the perturbation eigenvalue ratio is clustered. In Fig. 2, the vertical axis is the value of perturbation eigenvalue ratios, and the horizontal axis is the number of eigenvalues. The number of eigenvalues selected represents the number of factors in the final statistical factor. When $\gamma = 0.035$,[2] Fig. 2 shows that the ratio of

[2] The value of γ is not affected by the limited sample size. Specifically, there are a total of 8640 perturbed eigenvalues each year. However, Fig. 2 shows that the perturbed eigenvalues converge to 1 for all years after the 6th eigenvalue. The sample size of 8640, relative to 6, exhibits a substantial level of large-scale representation.

Table 1
(N,M) Values for unbalanced panels.

Year	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
N	977	927	474	939	452	1538	1216	1816	1558	1281	1306
M	8640	8640	8640	8640	8640	8640	8640	8640	8640	8640	8640

Note: N represents the number of stocks in the cross-section and M represents the number of 5-min return data on the time series. Compared to other years, the number of stocks in 2012 and 2014 is relatively small, mainly due to two reasons: firstly, in the raw data downloaded from the RESSET database, there are more stocks suspended during the day in 2012 and 2014, and after excluding this data, the number of stocks decreases; Secondly, in the 2015 bull market, the number of stocks in the market is surged, so the number of stocks in the market after 2015 is greater than before.

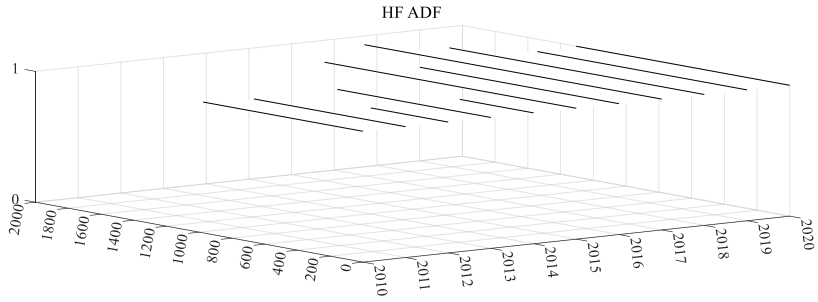

Fig. 1. Unit root test. Note: The X-axis represents the year, the Y-axis represents each stock, and the Z-axis represents whether the stock's return data for that year is stable. This paper conducts unit root tests on the annual return data of each stock according to the (N, M) data structure shown in Table 1. If the return sequence data of the Nth stock in year t is stable, it is 1, otherwise it is 0.

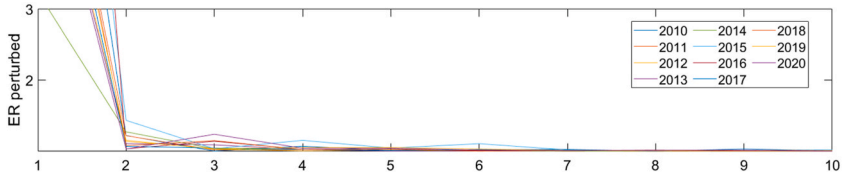

Fig. 2. Number of statistical factors.

perturbed eigenvalues after the 6th eigenvalue tends to stabilize for all years and clusters around 1 without any abnormal changes. Therefore, the first six factors are selected for preliminary model construction. By using the eigenvalue ratio method (Ahn & Horenstein, 2013; Onatski, 2010), the minimum risk factors can explain the maximum returns, thereby avoiding overfitting caused by principal component redundancy.

Fig. 2 shows that the first value of perturbation eigenvalue ratios for all years is larger than 3. In the PCA method, a higher eigenvalue signifies a greater amount of original information encapsulated within its corresponding eigenvector. Consequently, the statistical factor derived from the first eigenvector serves as the primary risk factor in the model. However, starting from the second eigenvalue, the ratio of eigenvalues exhibits variation across different years. Specifically, there are a total of 6 statistical systematic risk factors in 2015, after which the eigenvalue ratios converge to 1. The number of statistical risk factors is less than 6 for all the other years. To clearly distinguish the number of factors in each year, Fig. 3 shows the value of perturbation eigenvalue ratios for each year.

In Fig. 3, the number of risk factors varies across different years. Specifically, only two risk factors are evident in 2011–2012, while both 2013–2014 and 2016–2018 exhibit three risk factors. In 2010, four risk factors are observed, while 2019–2020 requires five factors, and 2015 necessitates six factors. These findings suggest that the optimal number of factors for most years is three, but certain years, such as 2019–2020 and 2015, require additional factors to better elucidate the relationship between stock returns and risk. Combined with historical stock market volatility, it is concluded that risk factors are time-varying in the cross-section, and higher volatility necessitates the inclusion of a greater number of risk factors to effectively capture stock market pricing behavior.

4.3. Statistical factor estimation

Through HF statistical factors, the intraday effects of factor return are additionally illustrated in Fig. 4, which displays the fundamental characteristics of the returns of the first three statistical factors at different time periods, respectively. And each variable represents the weighted average value corresponding to the respective time period on the x-axis, ranging from January 1, 2010, to December 31, 2020. As can be seen from Fig. 4, the analysis of the time period dimension reveals that stock returns tend to be higher at

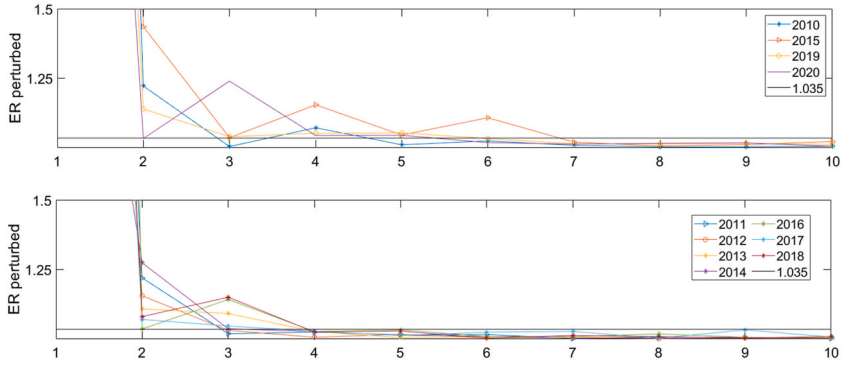

Fig. 3. Number of statistical factors by year.

the opening of the trading day, indicating the presence of a significant opening jump effect. Although the afternoon opening does not exhibit returns as high as those of the morning opening, it still represents a period of relatively elevated returns within the day. Furthermore, upon examining the different factor dimensions, it becomes apparent that the first HF factor exhibits greater sensitivity in reflecting stock market returns compared with the subsequent factors. The second HF factor demonstrates the secondary level of sensitivity, while the third HF factor exhibits the weakest relationship with stock market returns. Therefore, the first HF factor plays a critical role when incorporating HF factors to explain stock market risk premiums.

Table 2 reports the statistical factor characteristics for 2010–2020, including the contribution of each factor in explaining stock returns and the volatility degree of the factors. Based on the percentage of explanation, it is clear that the number of statistical factors with a non-zero percentage of explanation per year aligns closely with the total number of statistical factors per year. Furthermore, it is noteworthy that the first statistical factor consistently exhibits the highest percentage of explanation across the years. Specifically, in 2015 and 2016, the first factor's percentage of explanation reaches remarkable levels of 35 % and 36 %, respectively. This signifies that the first statistical factor plays a dominant role in influencing stock market returns during the period of 2015–2016. Moreover, the second statistical factor demonstrates a significantly larger percentage of explanation in 2014–2015 compared with other years, indicating that stock market returns in 2014–2015 are more responsive to the variations captured by the second statistical factor. However, to comprehensively understand the specific underlying factors influencing the statistical factors from an economic standpoint, further investigation is warranted.

Comparing the volatility degree of the statistical factor reveals that the volatility degree of the first statistical factor is from strong to weak during the period of 2010–2014. In 2015, there is a sudden and unusually strong volatility in the first statistical factor, followed by a subsequent return to a strong-to-gradually-weak volatility pattern from 2016 to 2020. The result aligns well with the historical trend, particularly considering the stock market's rapid surge to 5000 points in 2015. These findings indicate that constructing risk factors at the HF level is more accurate in reflecting stock market volatility, thus better portraying the marginal changes in returns and improving the pricing ability of stock market assets.

4.4. Economic interpretation of statistical factors

This paper takes the maximum combination weight values of the first 5 % of statistical factors 1–6 and classifies them according to industries, and the results are shown in Fig. 5. From Fig. 5, it can be seen that the combination weights of the first statistical factor are more consistent in different industries in both size and direction, so the maximum combination weight of the top 5 % of statistical factors comes from the whole market. The largest portfolio weights for the top 5 % of the second statistical factor come from the mining (B), manufacturing (C), information transmission, software and information technology services (I), and financial sectors (J). The third statistical factor has a maximum portfolio weight of the top 5 % from manufacturing (C), information transmission, software and information technology services (I), finance (J), and real estate (K). The fourth statistical factor with the largest combined weight of the top 5 % is derived from the mining (B) and finance sectors (J). The fifth statistical factor with the largest combined weight of the top 5 % comes from agriculture, forestry, animal husbandry and fishery (A), manufacturing (C), and finance (J). The sixth statistical factor has the largest combined weight of the top 5 % from manufacturing (C) and information transmission, software, and information technology services (I). The results help to assign economic meaning to the statistical factors.

Based on the analysis depicted in Fig. 5, we construct seven industry factors: market portfolio (equal-weighted market returns), manufacturing (equal-weighted manufacturing industry returns), finance (equal-weighted financial industry returns), information (equal-weighted returns from information transmission, software, and information technology services industry), energy (equal-weighted mining industry returns), real estate (equal-weighted real estate industry returns), and agriculture (equal-weighted returns from agriculture, forestry, livestock, and fishery sectors). Since there are differences in the sources of risk reflected by different risk factors in different years, the risk factors are economically explained by year.

Table 3 illustrates the broad correlations between the HF risk factors and industry factors by year. Since only two factors are available in 2011–2012 and the 3rd factors exist in all remaining years, the industry factors with the highest correlations with the first two and three risk factors are mainly explored. As can be seen from Fig. 5, the combination weights of the first two factors in different

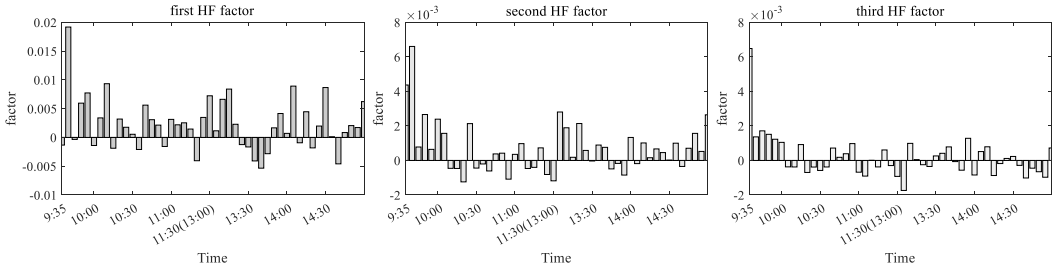

Fig. 4. Mean intraday returns of statistical factors.

Table 2
Statistical factor characteristics.

	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
Percentage of stock returns explained by statistical factors											
Factor 1	0.27	0.22	0.24	0.18	0.15	0.35	0.36	0.14	0.17	0.19	0.17
Factor 2	0.01	0.01	0.01	0.01	0.02	0.02	0.01	0.01	0.01	0.01	0.01
Factor 3	0.01	0.00	0.00	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Factor 4	0.01	0.00	0.00	0.00	0.01	0.01	0.00	0.01	0.00	0.01	0.01
Factor 5	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.01
Factor 6	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.00
The degree of volatility of the first three statistical factors											
Factor 1	0.18	0.15	0.11	0.14	0.09	0.25	0.22	0.17	0.18	0.17	0.16
Factor 2	0.03	0.03	0.03	0.03	0.03	0.06	0.03	0.04	0.04	0.04	0.05
Factor 3	0.02	0.02	0.02	0.03	0.02	0.04	0.03	0.04	0.04	0.04	0.04

Note: The explanatory ratio of statistical factors is calculated as $\frac{\lambda_{\widehat{K}t}}{\sum_{n=1}^{N}\lambda_{tn}}, t = (1,2,\ldots 11), \widehat{K} = (1,2,\ldots,6)$, N is the number of assets in year t, $\lambda_{\widehat{K}t}$ represents the $\widehat{K}$ th eigenvalue when the eigenvalues are ranked from the largest to the smallest after decomposition of the eigenvalues of the high-frequency return covariance matrix in year t, and $\sum_{n=1}^{N}\lambda_{tn}$ represents the sum of all eigenvalues in that year.

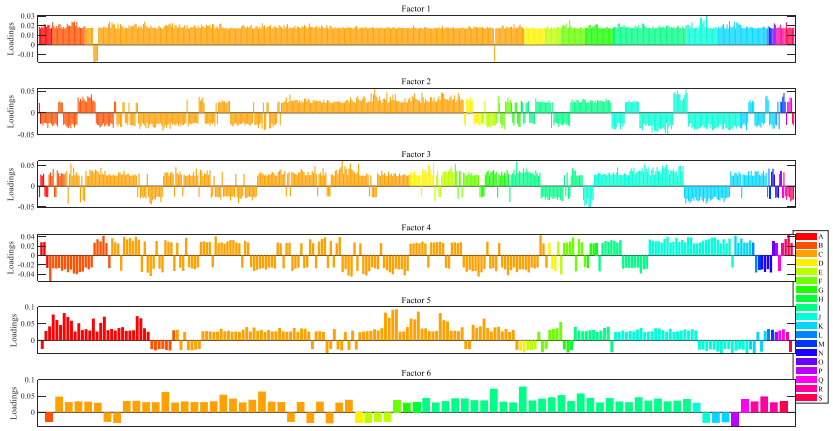

Fig. 5. Combined weights of HF statistical factors. Note: The industry is classified based on the 2012 CSRC industry classification standards.[3]

[3] A represents the agriculture, forestry, animal husbandry, and fishery industries; B represents the mining industry; C represents the manufacturing industry; D represents the power, heat, gas, and water production and supply industries; E represents the construction industry; F represents the wholesale and retail industries; G represents the transportation, warehousing, and postal industries; I represents the information transmission, software, and information technology service industries; J represents the financial industry; K represents the real estate industry; L represents the leasing and business services industries; M represents the scientific research and technology service industries; N represents the water conservancy, environment, and public facility management industries; Q represents the health and social work industries; R represents the cultural, sports, and entertainment industries; S represents integrated industries.

Table 3
GC between HF statistical factors and industry factors.

	GC	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
The first two HF statistical factors and industry factors												
F (M,C)	1.GC	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	2.GC	0.73	0.58	0.36	0.37	0.76	0.67	0.57	0.35	0.38	0.64	0.37
F (M,I)	1.GC	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	2.GC	0.29	0.48	0.34	0.69	0.46	0.55	0.65	0.53	0.78	0.75	0.73
F (M,J)	1.GC	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	2.GC	0.70	0.79	0.64	0.80	0.81	0.78	0.74	0.63	0.58	0.94	0.48
The first three HF statistical factors and industry factors												
F (M,J,C)	1.GC	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	2.GC	0.80	0.87	0.76	0.84	0.92	0.86	0.80	0.77	0.65	0.94	0.59
	3.GC	0.26	0.06	0.27	0.15	0.18	0.17	0.06	0.19	0.11	0.46	0.05
F (M,J,I)	1.GC	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	2.GC	0.72	0.80	0.64	0.81	0.81	0.81	0.75	0.73	0.81	0.95	0.77
	3.GC	0.03	0.42	0.33	0.64	0.42	0.46	0.65	0.30	0.52	0.75	0.38
Market	1.GC	0.95	0.97	0.30	0.94	0.95	0.85	0.90	0.94	0.95	0.97	0.92

Note: GC is the generalized correlation coefficient, 1. GC is the GC value between the HF risk factor and the principal component with the highest GC of the industry factor, and so on. F(M,C) is the market portfolio and manufacturing factor, F(M,I) is the market portfolio and information factor, and F (M,J) is the market portfolio and financial factor.

years are mainly derived from the market, manufacturing (C), finance (J), and information transmission, software, and information technology service (I) industries, where the combination weights of the first risk factor are shown to be derived from the market combination factor in different years. Therefore, the first industry factor is selected with equal weights from the market portfolio factor, and the second industry factor is selected from the manufacturing factor, the information factor, and the financial factor.

As can be seen from Table 3, the second generalized correlation coefficient of F (M, C) is relatively low around 0.5 in more than half of the years. Compared with the manufacturing factor, the GC coefficient of F (M, I) is higher in different years, but is not as high as that of F (M,J), which is above 0.7 in almost every year and lower only in 2020. It is seen that the first and second risk factors are highly correlated with the market portfolio factor and the financial factor. The third factor is chosen among the manufacturing factor and the information factor. Table 3 shows that the third GC coefficient of F (M, J, C) is below 0.2 in almost all years. In contrast, the third GC coefficient of F (M, J, I) is larger. In summary, the first three risk factors are mainly explained by the market portfolio factor, the financial factor, and the information factor. However, the correlation between F (M, J, I) in 2010 and 2020 is weaker, and the further analysis is needed due to the presence of four risk factors in 2010 and five risk factors in 2020.

Table 4 shows the industry portfolio factors with the largest GC coefficients corresponding to the top four risk factors in 2010, the top five risk factors in 2019–2020, and the top six risk factors in 2015. From Table 4, the industry portfolio factors corresponding to the four risk factors in 2010 are market portfolio factor, energy factor, real estate factor and financial factor. The industry portfolio factors corresponding to the five risk factors in 2019 and 2020 are market portfolio factor, information factor, manufacturing factor, financial factor and agricultural factor. And the industry portfolio factors corresponding to the six risk factors in 2015 are market portfolio factor, financial factor, information factor, manufacturing factor, agriculture factor, and energy factor. Although the industry factors corresponding to the first five risk factors in 2019 and 2020 are the same, there are differences in the GC coefficients. The first four GC coefficients in 2019 are above 0.7, and only the first three GC coefficients in 2020 are higher.

Combined with Tables 3 and it can be seen that the financial factor has weaker explanatory power for the risk factor in 2020, and the market portfolio factor, information factor and manufacturing factor have stronger explanatory power for the risk factor in 2020.2020 is the outbreak of COVID-19, so the pharmaceutical industry and manufacturing industry are more volatile in this year. The key point for the growth rate of manufacturing industry development from negative to positive is the advancement from traditional manufacturing to information technology, and then to digitalization. The deep application of digital technology has led to the rapid development of the information industry, so the manufacturing factor and the information factor became the main factors affecting the stock price.

5. HF risk factor model pricing capability test

5.1. Time-varying and constant coefficients

The stability of time-varying risk factor coefficients is crucial in stock market assets pricing. If the time-varying coefficients estimated from historical data are unstable, it will affect the effectiveness of asset value pricing. The stability of the time-varying risk factor coefficients mainly refers to the relationship between coefficients and time, based on the longitudinal comparison method. This method compares the coefficients calculated at different time and determine whether the values of the coefficients have changed over different periods, and the larger change indicates that the time-varying coefficients are more unstable. For daily data, the rolling window length of this paper is one month, i.e., the first 15 trading days of each month. For HF data, the rolling window length of this

Table 4
GC between HF statistical factors and industry factors.

	1.GC	2.GC	3.GC	4.GC	5.GC	6.GC
F (M,B,K,J, 2010)	1.00	0.89	0.82	0.32	0.00	0.00
F (M,I,C,J,A, 2019)	1.00	0.98	0.84	0.71	0.13	0.00
F (M,I,C,J,A, 2020)	1.00	0.88	0.82	0.31	0.21	0.00
F (M,J,I,C,A,B, 2015)	1.00	0.89	0.84	0.53	0.30	0.18

Note: The letters in parentheses represent the industry factors corresponding to the industry abbreviations, and the numbers represent the particular year. For example, F (M, B, K, J, 2010) is the market portfolio, energy, real estate and finance factors in 2010.

paper is720.[4] Then the time-varying risk factor coefficients are introduced, and the GC between the time-varying risk factor coefficients and the constant risk factor coefficients corresponding to different factors is estimated.[5] When the time-varying and corresponding constant coefficients have high correlation, the time-varying coefficients are more stable.

The results in Fig. 6 show that the GC results between the time-varying and constant coefficients of HF risk factors are stable, especially the first HF time-varying correlation with constant coefficients is close to 1, indicating that the time-varying coefficients estimated from the HF risk factors can effectively explain the stock market asset returns. In addition, the stability of the time-varying market factor coefficients as well as the time-varying industry portfolio factor coefficients are also high. However, the size factor and value factor in the Fama-French three-factor model, as well as the Fama-French five-factor model, exhibit significant volatility in the GC coefficients during different time periods. Furthermore, there is a weak correlation between the time-varying coefficients and the constant coefficients of the respective factors. It is seen that the time-varying coefficients calculated by HF are stable across time, while the time-varying coefficients of the Fama-French three and five factors at low frequencies are unstable. By comparing the GC results of daily and HF industry portfolio factors coefficients, and GC results of daily and HF market factors coefficients, it is found that the time-varying coefficients of both industry factors and market factors calculated by HF are more stable, which further indicates that the HF factor coefficients can be used as an effective estimate of the future risk factor coefficients and eventually achieve reasonable asset pricing.

5.2. Model fit optimization analysis

Based on the PCA method at the HF level, this paper further fits the daily risk factor and compares the explanatory power of high-frequency risk factors on stock market returns, the daily risk factor, the Fama-French three factors, the Fama-French five factors and the market factor, thus reflecting the pricing differences between different factor models. Table 5 shows the weighted average obtained from the regressions of different models on all stocks.

As can be seen from Table 5, under the HF perspective, the statistical factors model shows the largest improvement with 9 % compared with the CAPM model. In the daily perspective, the regression fit superiority ranking is time-varying Fama-French five-factor model $\succ$ time-varying statistical factor model $\succ$ time-varying Fama-French three-factor model $\succ$ time-varying CAPM model, but in the constant factor model, the regression fit superiority ranking is constant statistical factors model $\succ$ constant Fama-French three-factor model $\succ$ constant Fama-French five-factor model $\succ$ constant CAPM model. In general, the statistical factors model of HF fits better and has higher model pricing power.

5.3. Sharpe ratio

According to Barillas and Shanken (2018), the asset pricing capacity of each factor or factor model can be compared through the Sharpe ratios without introducing the test set assets. Fig. 7 reports the maximum Sharpe ratios of the different factors, and since the factors are based on the calculation of excess returns of traded assets, the risk premium can be calculated by using the time series mean. As shown in Fig. 7, the maximum portfolio annual Sharpe ratio of the factor is 2.15 for the HF statistical model based on three statistical factors, the annualized Sharpe ratio of the CAPM model constructed based on the HF market portfolio factor is 1.73, the combined annual Sharpe ratio of the three daily Fama-French factors is 1.52, and the combined annual Sharpe ratio of the five daily Fama-French factors is 1.72. The annualized Sharpe ratio of CAPM model based on daily market index factors is only 0.59. The confidence interval shows that the Sharpe ratio values of the above models are robust. The results above indicate that, compared to daily data, the factor model constructed based on HF data exhibits greater risk premium, which provides a better explanation for intraday fluctuations in stock prices and contains more information. In the HF perspective, the Sharpe ratio of statistical factor model is significantly higher than that of the CAPM model. And in the HF daily perspective, the Fama-French five-factor model has higher asset pricing power.

To further comprehend the pricing power of each factor in different factor models, this paper further constructs the risk premium of different factors based on the mean and covariance, i.e., the Sharpe ratio of factors. Since the statistical factors in this paper are

[4] Rolling on a monthly basis, there are 15 days per month, and 48 data per day, i.e. 720 H F data per month.

[5] The different factors mainly include the first three HF risk factors (3 H F Factors), the HF industry factors, the HF market factors, the daily Fama-French three factors, the daily industry factors, and the daily Fama-French 5 Factors. (Day Industry Factors), and day Fama-French five factors. In order to better compare the day Fama-French five factors with the HF factors, the first five HF risk factors (5 H F Factors) are further introduced.

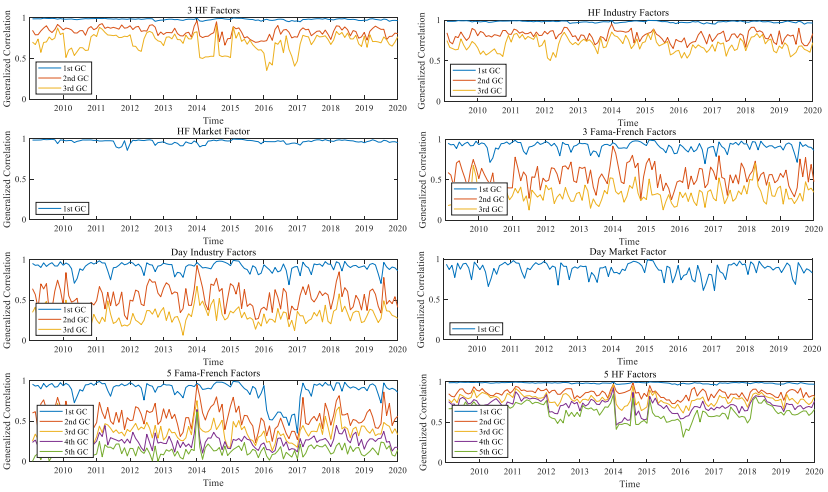

Fig. 6. GC between time variation and constant factor coefficients.

Table 5
Explanatory contribution of different factors to the return on assets.

	Statistical factors	Fama-French 3	Fama-French 5	Market
Time-varying HF R^2	0.25 (0.09)			0.16
Constant HF R^2	0.23 (0.08)			0.15
Time-varying daily R^2	0.50 (0.22)	0.47 (0.19)	0.53 (0.25)	0.28
Constant daily R^2	0.35 (0.13)	0.32 (0.10)	0.30 (0.08)	0.22

Note: The difference between R^2 of each factor model and R^2 of the corresponding CAPM model in parentheses in the table.

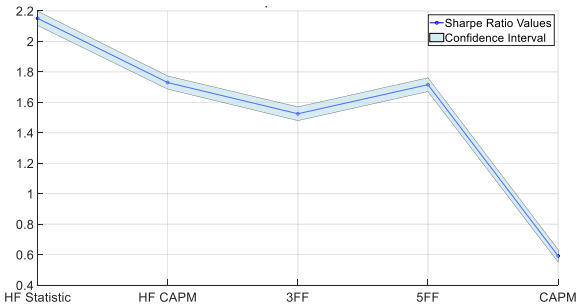

Fig. 7. Model maximum Sharpe ratio with Confidence Intervals. Note: The confidence interval calculated by the bootstrap standard errors of the Sharpe ratio for each factor model, where the bootstrap method is repeated a total of 1000 times.

calculated based on the excess asset returns, after normalizing all the factors using standard deviations, the time series averages are Sharpe ratios. As shown in Fig. 8, based on the frequency perspective, the Sharpe ratios are significantly higher than the daily industry and market factors for the HF statistical factor, industry factor and market factor. Based on the positive and negative perspective, most of the factors have positive Sharpe ratios and possess positive risk compensation, but the book-to-market ratio (HML) factor, the profitability (RMW) factor and the investment factor (CMA) factors have negative risk compensation. Overall, the statistical factor pricing ability of HF is much better than that of CAPM, Fama-French three factors and five factors.

6. Robustness test

6.1. factors, time-varying and constant coefficients and market liquidity indicators

Factor models of HF data are often ineffective due to their sensitivity to micro-structural noise, which largely depends on fluid interference. This paper examines the Pearson correlation between HF factors (the coefficients) and market liquidity indicators, with

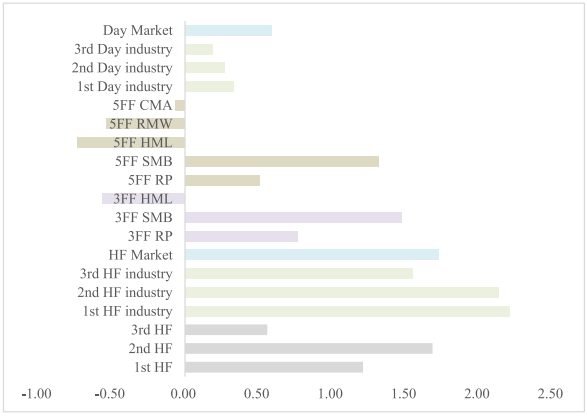

Fig. 8. Sharpe ratio of factors.

the results presented in Table 6. The correlation between the two depends on both the coefficient *r* and its significance. When the absolute value of *r* is less than 0.3, it is considered that the correlation between the two is "none or very weak". When the absolute value of *r* is between 0.3 and 0.5 and the coefficient is significant, the correlation between the two is weak. When the absolute value of *r* is greater than 0.5 and the coefficient is significant, the correlation between the two is strong (Shiha et al., 2020). From Tables 6 and it can be observed that the correlation coefficient between HF factors and market liquidity is close to 0, indicating that there is almost no correlation or a weak correlation. Furthermore, with the exception of the fifth beta, the absolute values of the correlations between the time-varying and constant betas and market liquidity are all less than 0.3, which can be considered as very weak correlations. Therefore, the HF factor constructed in this paper has weak correlation with micro-structure noise and is less affected by micro-structure noise.

6.2. *Out-of-sample test*

Out-of-sample test is essential to validate the predictive power and robustness of the model across different time periods and market conditions. This paper further introduces 5-min HF stock trading data from 9:30 on January 1, 2021 to 15:00 on December 31, 2021, and conducts out of sample test on HF statistical factor models, Fama-French three factor model, Fama-French five factor model, and daily CAPM model. This paper divides the HF stock data of 2021 into training and testing sets in a 7:3 ratio (Khagi et al., 2019), and the HF statistical factors can be calculated by using the training set data. By substituting HF statistical factors, Fama-French three factors, Fama-French five factors, and market factors into each corresponding model, the predicted return values of each stock at different time points in the test set can be calculated. The error is calculated by subtracting the predicted return values from the actual values in the test set. Table 7 reflects the mean absolute error (MAE) of all test set assets at all time points calculated by each model, which shows that the mean error calculated by the HF statistical factor model is the smallest.

6.3. *estimation results*

This paper further uses GMM to estimate the pricing errors and model-predicted returns for each model, specifically including the HF statistical factor model, HF CAPM model, Fama-French five-factor model, and Fama-French three-factor model. Fig. 9 shows the predicted and expected returns. As seen in Fig. 9, the estimated results of the HF statistical factor model indicate that the expected returns and predicted returns are nearly equal, with relatively small errors. The results suggest that the HF statistical factor model is more effective compared to the other models. This conclusion demonstrates the robustness of the results.

7. Conclusion

The advancement of information technology and the widespread adoption of the internet have propelled the era of big data. The dimensional catastrophe and frequency black box of factors have a significant impact on the pricing efficacy of factor models. For the Chinese stock market, previous research predominantly relies on index and daily data due to challenges in accessing HF data and limitations in model construction technology. As a result, there is a scarcity of studies that incorporate HF factor measures based on big data dimensions.

This paper employs data mining techniques to select 5-min data from 9:30 on January 1, 2010, to 15:00 on December 31, 2020, for all A-shares, visualizes 5-min HF statistical factor metrics by machine learning PCA model. The inner economic logic of HF statistical factors is elucidated based on the generalized correlation method. Additionally, this paper introduces both time-varying and constant versions of the CAPM, Fama-French three-factor, and five-factor models, which are analyzed with the HF statistical factor model for pricing power to reveal the contribution of statistical factors. The main conclusions are as follows: (1) Compared with the CAPM, Fama-French three and five factor models, the risk factors constructed based on the HF perspective improve the asset pricing model's

Table 6
Correlation between HF factors and market liquidity.

	HF factor1	HF factor2	HF factor3	HF factor4	HF factor5	HF factor6
all	0.050**	−0.002	0.055**	0.016	0.063***	−0.006
ShangHai	0.046**	−0.001	0.070***	0.031	0.062***	−0.008
ShenZhen	0.056**	−0.005	0.043*	−0.007	0.063***	−0.003
	Time beta1	Timebeta2	Timebeta3	Timebeta4	Time beta5	Time beta6
all	0.237**	0.277***	0.096***	0.220***	0.453*	0.190***
	Const beta1	Const beta2	Const beta3	Const beta4	Const beta5	Const beta6
all	0.237**	0.285***	−0.006	0.092***	0.477*	0.179***

Note: The market liquidity indicator is measured by daily turnover rate. Add up and average the data of 48 time periods of HF factors daily to obtain the factor sequence of HF factor daily frequency. Table 6 lists the correlation between the factor sequence of the first six HF factor daily frequency and the market liquidity of all exchanges, the Shanghai Stock Exchange and the Shenzhen Stock Exchange. Furthermore, Table 6 also reports the relationships between the time-varying and constant betas and market liquidity. Here, “Time beta" refers to the time-varying beta, while “Const beta" refers to the constant beta. Data are obtained from the RESSET Database.

Table 7
Mean absolute error for each model.

	HF statistic model	FF3 model	FF5 model	CAPM
MAE	$3.49*10^{-5}$	$162.88*10^{-5}$	$195.29*10^{-5}$	$232.75*10^{-5}$

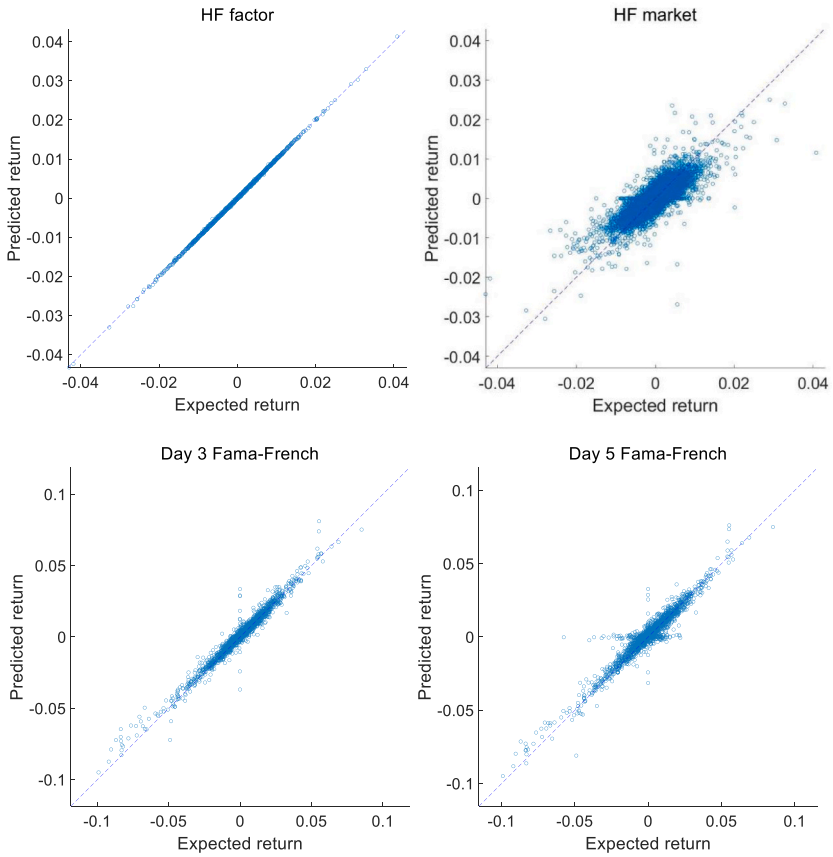

Fig. 9. Expected and predicted returns based on GMM.

explanatory power. Specifically manifested in three aspects: the time-varying beta of the HF factor model is more stable, the fit of the regression model is better, the annualized Sharpe ratio is the highest, and out-of-sample predicted values are very close to the true values. (2) The structure of HF factors is time-varying, and the more volatile the stock market, the higher the number of risk factors. (3) Since the first three HF statistical factors have the highest generalized correlations with the market portfolio factor, the financial factor,

and the information factor, the first three HF statistical factors can be explained by these three industry factors with practical significance.

The practical insight of this paper is that with the help of machine learning technology, HF information is of great value to the efficiency of stock market operation. (1) The stock market pricing mechanism should introduce HF data information and establish a database of HF data information. Because stock market transactions are HF existence with huge intraday transactions, through the real-time nature of HF data, the pricing process is updated in a timely manner to guarantee the stability of price fluctuations. (2) Stock market trading decisions are the result of a game of risk and return. In the consideration of stock market risk factors, HF information related to market portfolio, finance and information technology should be emphasized, so that the value of the enterprise itself can be more clearly defined. Accuracy in corporate value leads to high potential industries to attract more resources and thus achieve effective resource allocation. (3) The application of artificial intelligence in the processing of HF data in the stock market should be strengthened, using artificial intelligence methods to mine massive amounts of data and continuously optimize models. Machine learning improves the accuracy of interpreting stock returns based on identifying price trends, thereby reducing the cost and risk of trading.

Funding statement

This work was supported by the Humanities and Social Sciences Fund of the Ministry of Education, China (Grant No. 22YJA630047), the Young Scientists Fund of the National Natural Science Foundation of China (Grant No.72403217), and the National Natural Science Foundation of China (Grant No.72203104).

Conflict of interest statement

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

we would like to express our gratitude to the editors and anonymous reviewers for their valuable comments and suggestions. Any remaining errors are the sole responsibility of the authors.

Appendix

1. Factor Estimation

Suppose there are N assets that follow a stochastic process and each asset has a total of M HF data in time interval $[0,T]$. Then there are $M \times N$ dimensions of data. The time interval between each HF data is $\Delta M = T/M = t_{j+1} - t_j, j = 0,\ldots,M$. The initial price of stock is $\mathrm{p}(t_j)$, logarithmically recorded as $P(t_j) = \log(\mathrm{p}(t_j))$. To differentiate the variables in the approximate factor model $P(t_j) = \Lambda \mathrm{F}(t_j) + \mathrm{e}(t_j)$, the stochastic process of asset returns can be found, i.e., let $R_{j,i} = P_i(t_{j+1}) - P_i(t_j)$, $\Delta F_j = \mathrm{F}(t_{j+1}) - \mathrm{F}(t_j), \Delta e_{j,i} = e_i(t_{j+1}) - e_i(t_j)$. Thus, the approximate factor model can be written as the following equation:

$$\mathrm{R} = \Delta \mathrm{F} \Lambda^T + \Delta e \tag{1}$$

Where R is the log of HF return in dimension $\mathrm{M} \times \mathrm{N}$, ΔF is the risk factor in dimension $\mathrm{M} \times \mathrm{K}$, Λ^T is the factor loading matrix in dimension $\mathrm{K} \times \mathrm{N}$, and Δe is the asset-specific risk in dimension $\mathrm{M} \times \mathrm{N}$. R of the existing factor model is known in the estimation process, while the identification of the loadings Λ and risk factors ΔF need to be computed. According to Bai (2003), the estimation of $\widehat{\Lambda}$ and $\Delta\widehat{F}$ of the loadings Λ and risk factor ΔF are invertible transformations. In the approximate factor model, $\widehat{\Lambda}^T\widehat{\Lambda}/N = I_K$. $\Delta\widehat{F}^T\Delta\widehat{F}$ is assumed to be a diagonal matrix and there is a significant correlation between factor loadings and individuals. Specifically, there is a difference in the share of risk factors for different individuals, and the difference is the degree of influence of systematic risk factors on individuals. The traditional approximate factor model states that the systematic risk factor coefficients do not change over time, while the systematic risk factors change over time, i.e., the systematic risk factor vectors of asset returns vary over time. Express Eq. (1) in the form of estimation of loadings and risk factor as Eq. (2):

$$\mathrm{R} = \Delta\widehat{F}\widehat{\Lambda}^T \tag{2}$$

Squaring R yields

$$R^T R = \left(\Delta\widehat{F}\widehat{\Lambda}^T\right)^T\left(\Delta\widehat{F}\widehat{\Lambda}^T\right) = \widehat{\Lambda}\Delta\widehat{F}^T\Delta\widehat{F}\widehat{\Lambda}^T \tag{3}$$

$R^T R$ is an $\mathrm{N} \times \mathrm{N}$-dimensional square matrix, and the eigenvalue decomposition of $R^T R$ is performed. According to the definition of eigenvalue decomposition, for $R^T R = Q\Lambda Q^{-1}$, Q needs to be an $\mathrm{N} \times \mathrm{N}$-dimensional square matrix and Λ is a diagonal matrix. Therefore,

in Eq. (4), $\Delta\widehat{F}^T\Delta\widehat{F}$ is diagonal matrix. According to the assumptions of the approximate factor model, $\widehat{\Lambda}^T\widehat{\Lambda} \neq I_K$, $\widehat{\Lambda}^T\widehat{\Lambda}/N = I_K$. Let both sides of Eq. (3) be multiplied by $1/\sqrt{N}$, and then squared to obtain Eq. (4) as:

$$(1/N)R^TR = \frac{1}{N}\widehat{\Lambda}\Delta\widehat{F}^T\Delta\widehat{F}\widehat{\Lambda}^T = \left(\widehat{\Lambda}\Big/\sqrt{N}\right)\Delta\widehat{F}^T\Delta\widehat{F}\left(\widehat{\Lambda}^T\Big/\sqrt{N}\right) \tag{4}$$

Then, $\left(\widehat{\Lambda}^T/\sqrt{N}\right)\left(\widehat{\Lambda}/\sqrt{N}\right) = \widehat{\Lambda}^T\widehat{\Lambda}/N = I_K$. Therefore, according to the definition of eigenvalue, $\widehat{\Lambda}/\sqrt{N}$ is the eigenvariable of matrix $(1/N)R^TR$, and the load $\widehat{\Lambda}$ is the eigenvector of $(1/N)R^TR^*\sqrt{N} = (1/\sqrt{N})R^TR$. And $\Delta\widehat{F}^T\Delta\widehat{F}$ is the diagonal matrix composed of eigenvalues. According to the definition of the unsupervised dimensionality reduction method, the larger the eigenvalue, the larger its corresponding variance, representing more information contained, so about the load is equal to taking the eigenvector corresponding to the K largest eigenvalues. Then the statistical factor is:

$$\underbrace{\Delta\widehat{F}}_{M\times\widehat{K}} = \underbrace{R}_{M\times N}\underbrace{\widehat{\Lambda}}_{N\times\widehat{K}}\left(\widehat{\Lambda}^T\widehat{\Lambda}\right)^{-1} \tag{5}$$

It is important to note that $\frac{1}{N}R^TR$ is the quadratic covariance of asset prices conditional on finite assets N. And in equations (1)–(5), the observations of M and the number of assets N obey $M,N\to\infty$. This condition indicates that the values in the time series are close to continuous and the data are large samples in the cross-section, i.e., high-dimensional data, when the estimated $\Delta\widehat{F}$ and $\widehat{\Lambda}$ are consistent estimates of ΔF and Λ.

2. Constants and Time-Varying Betas Estimation

The quadratic covariance of the statistical risk factor is $\Delta\widehat{F}^T\Delta\widehat{F}$. In addition to calculating the quadratic covariance, $\Delta\widehat{F}$ can also be used to calculate the constant and time-varying factor loadings in the regression of high-frequency data. The loadings $\widehat{\Lambda}$ obtained from the eigenvalue decomposition of $\left(1/\sqrt{N}\right)R^TR$ are not time-varying, and in this paper, time-varying betas calculated from statistical factors are obtained in order to investigate whether they are more stable than time-varying betas calculated by linear regression of systematic risk factors in existing factor models. In estimating the time-varying beta, N-dimensional R can be represented by the following local statistical factor model:

$$R = \beta_{i,s}^T\Delta\widehat{F} + \Delta e, i = 1,\ldots,N, \mathrm{t}\in(T_s,T_{s+1}), s = 1,\ldots,S \tag{6}$$

where R is the M × N-dimensional stochastic wandering process representing the logarithmic return of asset N at moment M. $\Delta\widehat{F}$ is the $M\times\widehat{K}$-dimensional stochastic wandering process known as factor return. $\beta_{i,s}$ is the $N\times\widehat{K}$-dimensional vector representing the factor coefficients and $\beta_{i,s}$ represents the exposure to the systematic factor $\Delta\widehat{F}$. The residual e_i is the K × N-dimensional random wandering process and represents the idiosyncratic risk component in addition to the systematic risk. For the time interval of $[T_s,T_{s+1}]$, β_s is a constant. In contrast, in this paper, the time-varying factor loadings are calculated by rolling window regression based on previous time-varying betas with known excess returns and risk factors in the empirical test. The loadings estimated with different local time windows have different factors and residual structures. In this paper, we use 1 month as the time window for rolling regression to calculate time-varying betas. The time-varying betas are calculated in the same way as the constant betas for each month, but the number of betas per category is 12 per year and different every month, because the factors vary from month to month.

The above model estimates a time-varying factor model, and to simplify the model, the subscript s can be removed, resulting in a local factor model with estimated constant coefficients:

$$R = \beta_i^T\Delta\widehat{F} + \Delta e, i = 1,\ldots,N, \mathrm{t}\in[0,T] \tag{7}$$

Where the time span of the whole sample is T. For a fixed time interval $[0,T]$, the time period is partitioned into $t_0 = 0, t_1 = \Delta M, t_2 = 2\Delta M, t_3 = 3\Delta M,\ldots,t_M = M\Delta M$, where the time increment is $\Delta M = t_{j+1} - t_j = T/M$. For example, the sample period used in this paper is January 1, 2010, 9:30 - December 31, 2020, 15:00. Then T is December 31, 2020, 15:00. The HF data used in this paper is 5 min, so $\Delta M = 5$. Then M is the total number of 5-min data during the sample period. At this point, the constant beta for the whole sample period can be estimated as:

$$\underbrace{\beta}_{N\times\widehat{K}} = \underbrace{R^T}_{N\times M}\underbrace{\Delta\widehat{F}}_{M\times\widehat{K}}\left(\Delta\widehat{F}^T\Delta\widehat{F}\right)^{-1} \tag{8}$$

Where each asset has $\widehat{K}$ different constant beta, i.e., each systematic factor corresponds to a constant beta. Unlike the existing observable factor models, the statistical factors are based on the $\widehat{K}$ systematic risk factors obtained from PCA, and the most significant feature is that $\widehat{K}$ risk factors are orthogonal and uncorrelated with each other. Therefore, Eq. (8) can be used to calculate different constant betas and obtain consistent estimates.

3. Number of Factors

After estimating the risk factors F, factor loadings Λ and beta coefficients, it is necessary to determine the optimal number of factors $\widehat{K}$. Assuming a total number of K risk factors, modeling is required to select the optimal number of factors $\widehat{K}$ from K risk factors. According to the machine learning PCA method, it is known that the larger the eigenvalue, the rawer information the data contains. Therefore, the selection of the factor number is closely related to the eigenvalues, and the number of selected system factors is equivalent to the number of selected eigenvalues. According to Ahn and Horenstein (2013), the size of the eigenvalue ratio is considered when selecting the number of risk factors, and the eigenvalue ratio ER_k is the ratio between two consecutive eigenvalues. Since the eigenvalues are arranged in descending order, the eigenvalue ratio is greater than 1. The threshold $1+\gamma$ is a set for the size of the eigenvalue ratio ER_k. If $ER_k \succ 1+\gamma$, the amount of information contained meets the criteria, and the number of risk factors can be clearly determined. For the eigenvalue ratio, when the eigenvalues in the eigenvalue matrix $\Delta\widehat{F}^T\Delta\widehat{F}$ are very small, such as infinitely close to 0 or even equal to 0, the eigenvalue ratio is very large or even erroneous, and thus the results are biased. To avoid such problems, this paper uses the perturbation method to estimate eigenvalue ratio.

The perturbation method is a common method in matrix theory to transform a general matrix problem into an invertible matrix problem. Since the sufficient necessary condition for an invertible matrix is that none of the matrix determinants is 0, for the diagonal matrix $\Delta\widehat{F}^T\Delta\widehat{F}$, if it is transformed so that none of the determinants is 0, it means that there are no eigenvalues equal to 0. From the definition of matrix, we know that for any square matrix A, exiting $t \in R^+$ makes $tI_n + A$ an invertible matrix. If the invertible matrix of the matrix holds and is continuous with respect to t, the problem can be obtained independent of the values of t, i.e., it also holds when $t = 0$, thus obtaining that the problem holds for an irreversible matrix. Therefore, the results obtained by using the perturbation method do not affect the actual estimation results. Regarding the choice of t, there are infinitely many t satisfying the condition. In summary, this paper uses the perturbation method to find the eigenvalue ratios. At first, we assume that there is a total of $\widehat{K}$ factors selected in the end. Then the eigenvalues of R^TR can be found according to Eq. (3), and by arranging all the eigenvalues in order from the largest to the smallest, $\lambda_1 \succcurlyeq \lambda_2 \succcurlyeq \ldots \succcurlyeq \lambda_N$ can be obtained. The perturbation method is used to find the perturbed eigenvalues as follows:

$$\widehat{\lambda}_k = \lambda_k + g(N,M), (k = 1,2,\ldots,N) \tag{9}$$

Where $g(N,M) \in R^+$, with reference to Pelger (2019), the simulation of $g(N,M)$ value is:

$$g(N,M) = \sqrt{N}{*}median\{\lambda_1,\lambda_2,\ldots,\lambda_N\} \tag{10}$$

The final perturbation eigenvalues used to calculate the eigenvalue ratios are as follows:

$$\widehat{\lambda}_k = \lambda_k + \sqrt{N}{*}median\{\lambda_1,\lambda_2,\ldots,\lambda_N\}, (k = 1,2,\ldots,N) \tag{11}$$

According to equation (11), the perturbation eigenvalue ratio statistic can be calculated as:

$$ER_k = \widehat{\lambda}_k/\widehat{\lambda}_{k+1}, (k = 1,2,\ldots,N-1) \tag{12}$$

At this point, the perturbation eigenvalues are arranged in descending order as $\widehat{\lambda}_1 \succcurlyeq \widehat{\lambda}_2 \succcurlyeq \ldots \succcurlyeq \widehat{\lambda}_N$ to calculate ER_k, and the following result is obtained:

$$\widehat{K}(\gamma) = max\{k \preccurlyeq N-1 : ER_k \succ 1+\gamma\}, \gamma \succ 0 \tag{13}$$

Eq. (13) is the maximum value of k that is selected to satisfy the condition of $ER_k \succ 1+\gamma$.

Data availability

Data will be made available on request.

References

Ahn, S. C., & Horenstein, A. R. (2013). Eigenvalue ratio test for the number of factors. *Econometrica, 81*(3), 1203–1227.
Andersen, T. G., Riva, R., & Thyrsgaard, M. (2023). Intraday cross-sectional distributions of systematic risk. *Journal of Econometrics, 235*(2), 1394–1418.
Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica, 71*(1), 135–171.
Barillas, F., & Shanken, J. (2018). Comparing asset pricing models. *The Journal of Finance, 73*, 715–754.
Barinov, A. (2014). Turnover: Liquidity or uncertainty? *Management Science, 60*(10), 2478–2495.
Beber, A., Brandt, M. W., & Luisi, M. (2015). Distilling the macroeconomic news flow. *Journal of Financial Economics, 117*(3), 489–507.
Bekaert, G., & Hoerova, M. (2014). The vix, the variance premium and stock market volatility. *Journal of Econometrics, 183*(2), 181–192.
Bergbrant, M. C., & Kelly, P. J. (2016). Macroeconomic expectations and the size, value, and momentum factors. *Financial Management, 45*(4), 809–844.
Brunnermeier, M. K., Sockin, M., & Xiong, W. (2022). China's model of managing the financial system. *The Review of Economic Studies, 89*(6), 3115–3153.
Cartea, A., & Jaimungal, S. (2013). Modelling asset prices for algorithmic and high-frequency trading. *Applied Mathematical Finance, 20*(6), 512–547.
Cayirli, O., Kayalidere, K., & Aktas, H. (2022). Asset pricing in a multifactor setting. *Borsa Istanbul Review, 22*(6), 1062–1068.

Chamberlain, G., & Rothschild, M. (1983). Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica, 51*(5), 1281–1304.
Chen, L., Pelger, M., & Zhu, J. (2023). Deep learning in asset pricing. *Management Science*, (2), 1–37.
Cochrane, J. H. (2011). Presidential address: Discount rates. *The Journal of Finance, 66*(4), 1047–1108.
Fama, E. F., & French, K. R. (2020). Comparing cross-section and time-series factor models. *Review of Financial Studies, 33*(5), 1891–1926.
Fama, E. F., French, K. R., & Booth, D. G. (1993). Differences in the risks and returns of NYSE and NASD stocks. *Financial Analysis Journal, 49*(1), 37–41.
Grillini, S., Ozkan, A., & Sharma, A. (2019). Pricing of time-varying illiquidity within the Eurozone: Evidence using a Markov switching liquidity-adjusted capital asset pricing model. *International Review of Financial Analysis, 64*, 145–158.
Guobužaitė, R., & Teresienė, D. (2021). Can economic factors improve momentum trading strategies? The case of managed futures during the COVID-19 pandemic. *Economies, 9*(2), 86.
Hendershott, T., Menkveld, A. J., & Praz, R. (2022). Asset price dynamics with limited attention. *Review of Financial Studies, 35*(2), 962–1008.
Herskovic, B., Kind, T., & Kung, H. (2023). Micro uncertainty and asset prices. *Journal of Financial Economics, 149*(1), 27–51.
Hollstein, F., Prokopczuk, M., & Wese, S. C. (2020). The conditional capital asset pricing model revisited: Evidence from high-frequency betas. *Management Science, 66*(6), 2474–2494.
Jenter, D., & Lewellen, K. (2015). CEO preferences and acquisitions. *The Journal of Finance, 70*(6), 2813–2852.
Kelly, B. T., Moskowitz, T. J., & Pruitt, S. (2021). Understanding momentum and reversal. *Journal of Financial Economics, 140*(3), 726–743.
Khagi, B., Kwon, G. R., & Lama, R. (2019). Comparative analysis of Alzheimer's disease classification by CDR level using CNN, feature selection, and machine-learning techniques. *International Journal of Imaging Systems and Technology, 29*(3), 297–310.
Kozak, S., Nagel, S., & Santosh, S. (2020). Shrinking the cross-section. *Journal of Financial Economics, 135*(2), 271–292.
Kritzman, M., Li, Y., Page, S., & Rigobon, R. (2011). Principal components as a measure of systemic risk. *Journal of Portfolio Management, 37*(4), 112–126.
Leippold, M., Wang, Q., & Zhou, W. (2022). Machine learning in the Chinese stock market. *Journal of Financial Economics, 145*(2), 64–82.
Lettau, M., & Pelger, M. (2020). Factors that fit the time series and cross-section of stock returns. *Review of Financial Studies, 33*(5), 2274–2325.
Li, Z., & Rao, X. (2022). Evaluating asset pricing models: A revised factor model for China. *Economic Modelling, 116*, Article 106001.
Louzis, D. P., Xanthopoulos-Sisinis, S., & Refenes, A. P. (2013). The role of high-frequency intra-daily data, daily range and implied volatility in multi-period value-at-risk forecasting. *Journal of Forecasting, 32*(6), 561–576.
Ma, T., Leong, W. J., & Jiang, F. (2023). A latent factor model for the Chinese stock market. *International Review of Financial Analysis, 87*, Article 102555.
Ma, M., Zheng, L., & Yang, J. (2021). A novel improved trigonometric neural network algorithm for solving price-dividend functions of continuous time one-dimensional asset-pricing models. *Neurocomputing, 435*, 151–161.
Mallinger-Dogan, M., & Szigety, M. C. (2014). Higher-frequency analysis of low-frequency data. *Journal of Portfolio Management, 41*(1), 121–138.
Novy-Marx, R. (2013). The other side of value: The gross profitability premium. *Journal of Financial Economics, 108*(1), 1–28.
Onatski, A. (2010). Determining the number of factors from empirical distribution of eigenvalues. *The Review of Economics and Statistics, 92*(4), 1004–1016.
Pástor, Ľ., Stambaugh, R. F., & Taylor, L. A. (2022). Dissecting green returns. *Journal of Financial Economics, 146*(2), 403–424.
Pelger, M. (2019). Large-dimensional factor modeling based on high-frequency observations. *Journal of Econometrics, 208*(1), 23–42.
Pelger, M. (2020). Understanding systematic risk: A high-frequency approach. *The Journal of Finance, 75*(4), 2179–2220.
Qiao, K., & Dam, L. (2020). The overnight return puzzle and the "T+1" trading rule in Chinese stock markets. *Journal of Financial Markets, 50*, Article 100534.
Renault, T. (2017). Intraday online investor sentiment and return patterns in the US stock market. *Journal of Banking & Finance, 84*, 25–40.
Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The Journal of Finance, 19*(3), 425–442.
Shiha, A., Dorra, E. M., & Nassar, K. (2020). Neural networks model for prediction of construction material prices in Egypt using macroeconomic indicators. *Journal of Construction Engineering and Management, 146*(3), Article 04020010.
Wang, C. D., Chen, Z., Lian, Y., & Chen, M. (2022). Asset selection based on high frequency Sharpe ratio. *Journal of Econometrics, 227*(1), 168–188.
Yao, H., Xia, S., & Liu, H. (2022). Six-factor asset pricing and portfolio investment via deep learning: Evidence from Chinese stock market. *Pacific-Basin Finance Journal, 76*, Article 101886.
Zhang, C., Liu, Z., & Liu, Q. (2021). Jumps at ultra-high frequency: Evidence from the Chinese stock market. *Pacific-Basin Finance Journal, 68*, Article 101420.
Zhu, X., Zhang, H., & Zhong, M. (2017). Volatility forecasting using high frequency data: The role of after-hours information and leverage effects. *Resources Policy, 54*, 58–70.