Study of Stock Return Prediction Based on Big Data

Shiqi Chen [1, *]
[*]shiqichen0617@gmail.com

Wardlaw Hartridge School

Abstract. The stock market is one of the most important components of the capital market, and thus identifying the reasons for changes in expected stock returns is key to building a well-functioning capital market. Based on Gordon's dividend growth model, this paper composes company fundamental indicators from five aspects: profitability, growth, corporate governance, potential value and safety, and obtains a company characteristics data set containing 115 indicators. Principal component analysis, Enigma Macbeth regression method, predictive portfolio method, composite principal component analysis and partial least squares method are applied to reduce the dimensionality of the above data set and construct a comprehensive quality index of company fundamentals respectively. The quality index Q^{PLS} constructed based on the partial least squares method is found to have the strongest and relatively stable predictive power through the uni variate portfolio analysis method. This paper confirms the importance of quantitative big data integration in stock market research and enriches the research on the frontier of "big data + asset pricing".

Keywords. fundamentals; big data; stock return forecasting.

1 Introduction

Specific investment strategies developed in response to fundamentals have historically had higher returns than those adjusted for systematic risk [2]. For example, investing in value stocks can yield higher returns than growth stocks, known as the value effect [1]. The effort to categorize and analyze these vast stock market anomalies has taken more than two decades to expand from the three-factor asset pricing model [1] to the current four-factor [3] and five-factor models [4]. Although these two competing models have different theoretical foundations, they both consider the pricing factors of earnings and investment, which shows the importance of the earnings and investment categories of stock market anomalies.

The research problem in this paper is based on a large number of fundamental indicators, including earnings and investment, to construct indicators that measure the comprehensive quality of fundamentals and to test their predictive power of cross-sectional returns of Chinese stocks to prove the effectiveness of fundamental analysis in the Chinese stock market. The first question to be addressed is which indicators are the underlying indicators and how to integrate the large number of indicators. We will use Gordon's dividend growth model and the industry's investment experience to summarize fundamental indicators in five dimensions: profitability, growth capability, corporate governance, potential value, and safety; and we will use principal component analysis, Enigma-Macbeth regression, forecast portfolio analysis, principal component composite analysis, and partial least squares to construct comprehensive

MSIEID 2022, December 09-11, Chongqing, People's Republic of China

DOI 10.4108/eai.9-12-2022.2327593

quality indicators of fundamentals, and through stock return prediction ability comparison, select the optimal parameter as the final proxy variable.

2 Data and sample

Financial data, monthly and daily stock returns, Fama & French (1993) factors, and risk-free returns for China are obtained from the CSMAR database. The sample includes all A-shares on the Shanghai Stock Exchange and Shenzhen Stock Exchange, and the interval is from January 1999 to December 2021.

3 Integrated Quality Factor Predictive Power analysis

It is first assumed that the composite quality factor in period t explains the excess stock returns in period t+1, and the two are linearly related.

$$E_t(R_{t+1}) = \alpha + \beta Q_t \tag{1}$$

where.$Q_t$denotes the composite quality of the firm as reflected by the fundamentals. And the actual return to investors in period t+1 is the sum of conditional returns and stochastic volatility (equation 2-2), whereε_{t+1} is the stochastic volatility term, which is unobservable andQ_t uncorrelated.

$$R_{t+1} = E_t(R_{t+1}) + \varepsilon_{t+1} = \alpha + \beta Q_t + \varepsilon_{t+1} \tag{2}$$

Let$X_t = (x_{1,t}, x_{2,t}, \dots, x_{N,t})'$ be the set of N × 1-dimensional firm fundamental indicators in period t. In this section, 115 fundamental indicators in five dimensions of profitability, growth, governance strategy and value are selected, i.e. N=115. These indicators characterize different aspects of the firm and may have different units of measure. Therefore, it is necessary to standardize each indicator separately, i.e., assume that any of the characteristic variables$x_{i,t} = (i = 1, \dots, N)$ has a mean of 0 and a variance of 1 in the cross-section. inQ_t the premise that it is the only predictor of the stock's future returns, each characteristic variable$x_{i,t}$ has the following factor structure:

$$x_{i,t} = \eta_{i.0} + \eta_{i.1}\left(E_t(R_{t+1}) - \overline{E_t(R_{t+1})}\right) + e_{i,t}, i = 1, \dots, N \tag{3}$$

where$\eta_{i.1}$ is the sensitivity of the ith characteristic variable to the expected stock return, and$\overline{E_t(R_{t+1})}$ is the cross-sectional mean of the expected stock returns, and$e_{i,t}$ is the individual error associated with the ith characteristic variable only. the PLS method can effectively filter the irrelevant common errors, which precisely overcomes the shortcomings of PCA. Thus, the ultimate goal of the PLS method is to extract the common factor with the largest covariance with stock expected returns from a series of predictors for the purpose of dimensionality reduction, otherwise these large number of predictors would cause the model to fall into a high-dimensional trap. Compared to methods such as PCA, the factors constructed based on PLS may not be the most important public elements of the predictors, but have the strongest predictive power.

For this purpose, PLS can be implemented by two least squares (OLS) methods. First, in the cross-section, the t period stock returns areR_t on each lagged one-period fundamental indicator$x_{i,t-1}$ Regressions are conducted on.

$$R_t = \pi_{i,0} + \pi_i x_{i,t-1} + u_{i,t-1}, t = 1, \dots, N \tag{4}$$

In equation (2-4), theπ_i denotes the firm characteristics$x_{i,t-1}$ on the expected return on the stockR_t and can be viewed as the sensitivity of$\eta_{i.1}$ a proxy variable for the Since stock returnsR_t predictable component is caused by theQ_{t-1} by, the fundamental indicators are correlated with the predictable part of stock returns and not with the unpredictable part, thereforeπ_i describes the contribution of each firm characteristic to the composite quality index.

In the second step, on the time series, run each fundamental indicator$x_{i,t-1}$ on theπ_i linear regression of the estimates.

$$x_{i,t-1} = \omega_t + Q_t^{PLS}\widehat{\pi}_\iota + v_{i,t}, i = 1, \dots, N \tag{5}$$

In summary, the PLS method organically combines equations (3) and (4) to obtain the calibrated indices. The first step of the cross-sectional regression, first obtains$\widehat{\pi}_\iota$, the contribution of each firm characteristic to the final index; the second step of the cross-sectional regression regresses each firm characteristic on this contribution to obtain the final calibrated index. the PLS method uses t+1 period stock returns to downscale the fundamentals, sieve out individual noise$e_{i,t}$ and thus extracts the fraction of returns that can be predictedQ_t.[6]

4 Integrated Quality Factor Predictive Power analysis

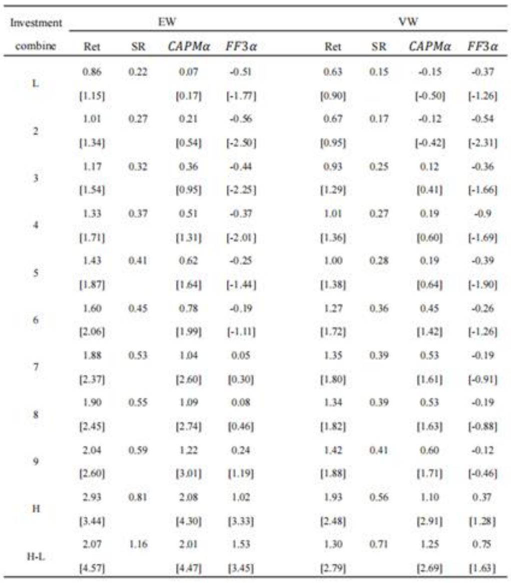
4.1 Comparative analysis of the predictive power of different methods for constructing composite quality factors

Uni-variate portfolio analysis was applied to compare the composite quality factors constructed by different methodsQ^{PLS}, Q^{PCA}, Q^{FM}, Q^{FC} and Q^{CPCA} for their predictive power. At the beginning of each month, stocks are divided into 10 groups based on the previous month's composite quality factor, and portfolios are constructed with both equal weights and market capitalization weights, while long-short hedged portfolios are constructed by buying high-quality and selling low-quality stocks, and the average monthly raw return of each portfolio is calculated and reconstructed at the beginning of the following month. Then, based on the CAPM and the Fama & French (1993) three-factor model, the excess returns of the portfolios are calculated separately $CAPM\alpha$ and $FF3\alpha$. When the portfolios are constructed with equal weights, the composite quality factors constructed by the four methods, except for the principal component analysis, have some degree of predictive power. For example, The average monthly returns of the hedged portfolios constructed by Q^{PLS}, and Q^{FM}, Q^{FC} and Q^{CPCA} are 2.07% (t=4.57), 1.06% (t=3.32), 0.84% (t=1.95), and 0.76% (t=1.67), respectively, here the investment strategies with quality factors constructed by partial least squares and Enigma Macbeth regressions are significant at the 1% level. When controlling for the market factor, the predictive power of the quality factor constructed by the composite principal component analysis is not significant; after continuing to control for the size and value factors, onlyQ^{PLS},theQ^{FM}'s hedged portfolio's excess returns remain significant at 1.53% (t=3.45), 0.91% (t=2.78), respectively, regardless of the size and salience of $FF3\alpha$ the investment strategy Q^{PLS} performs better. In practice, it is more feasible to construct a portfolio operation with market capitalization weighting. It turns out that at this point onlyQ^{PLS}has predictive power for cross-sectional stock returns. Investors can earn an average initial return of 1.30% per month (t=2.79) by buying high quality and selling low quality stocks and an excess return of 1.25% (t=2.69) under the CAPM model, both significant at the 1% level.[7]

4.2 Predictive power analysis of integrated quality factors based on partial least squares construction

The comparison reveals thatQ^{PLS} the strongest predictive power, Figure 1 reports in detail theQ^{PLS} results of the uni-variate portfolio analysis. For the equally weighted constructed portfolios, it can be observed that the monthly average raw return of the stock increases monotonically from 0.86% (t=1.15) to 2.93% (t=3.44) as the quality of the firm increases, with an annual difference of nearly 25% between the two. When controlling for the market factor, the portfolios' returns still show a monotonically increasing pattern, with the hedged portfolio's return of 2.01% (t=4.47) not far from the original return of 2.07% (t=4.57). When controlling for the Fama & French (1993) three factors, the return on the hedged portfolio drops to 1.53% (t=3.45), at which point the contribution of the long side is nearly 67% (1.02%/1.53%), indicating that the contribution of the long side basedQ^{PLS} investment strategy mainly from the long side, which is less affected by short selling restrictions and is more viable in the Chinese equity market. The Sharpe ratio also increases withQ^{PLS} increase, suggesting that the higher the quality of the company, the greater the compensation per unit of risk given to investors, and the more popular it should be with mean-variance investors.[8]

When the portfolio is constructed with market capitalization weighting, the raw return, excess return and Sharpe ratio show similar trend changes, but all are smaller than the corresponding values for the equally weighted constructed portfolio. This is most likely due to the fact that smaller market capitalization stocks contribute more than larger stocks toQ^{PLS} predictive power, then constructing the portfolio with market capitalization weights relatively reduces the weight of small stocks, which not only reduces the return of the equal quality stock portfolio, but also reduces the difference in returns of extreme quality company stocks, i.e., the return of the hedged portfolio becomes smaller. Further tests will follow to examine the relationship between firm size andQ^{PLS} the relationship between predictive power. It is also found that when controlling for the market, size and value factors, the $FF3\alpha$ of the hedged portfolio is an average of 0.75% per month (t=1.63), which is insignificant but less different than the t-statistic at the 10% significance level. Also at this point, the lowQ^{PLS}company's stock excess return is -0.37%, while the highQ^{PLS} stock excess return is 0.37%, indicating that the long and short sides contribute similarly to the return of the hedged portfolio, also reflecting a difference from the equal market capitalization weighted case.[9]



Investment	EW				VW			
combine	Ret	SR	$CAPM\alpha$	$FF3\alpha$	Ret	SR	$CAPM\alpha$	$FF3\alpha$
L	0.86	0.22	0.07	-0.51	0.63	0.15	-0.15	-0.37
	[1.15]		[0.17]	[-1.77]	[0.90]		[-0.50]	[-1.26]
2	1.01	0.27	0.21	-0.56	0.67	0.17	-0.12	-0.54
	[1.34]		[0.54]	[-2.50]	[0.95]		[-0.42]	[-2.31]
3	1.17	0.32	0.36	-0.44	0.93	0.25	0.12	-0.36
	[1.54]		[0.95]	[-2.25]	[1.29]		[0.41]	[-1.66]
4	1.33	0.37	0.51	-0.37	1.01	0.27	0.19	-0.9
	[1.71]		[1.31]	[-2.01]	[1.36]		[0.60]	[-1.69]
5	1.43	0.41	0.62	-0.25	1.00	0.28	0.19	-0.39
	[1.87]		[1.64]	[-1.44]	[1.38]		[0.64]	[-1.90]
6	1.60	0.45	0.78	-0.19	1.27	0.36	0.45	-0.26
	[2.06]		[1.99]	[-1.11]	[1.72]		[1.42]	[-1.26]
7	1.88	0.53	1.04	0.05	1.35	0.39	0.53	-0.19
	[2.37]		[2.60]	[0.30]	[1.80]		[1.61]	[-0.91]
8	1.90	0.55	1.09	0.08	1.34	0.39	0.53	-0.19
	[2.45]		[2.74]	[0.46]	[1.82]		[1.63]	[-0.88]
9	2.04	0.59	1.22	0.24	1.42	0.41	0.60	-0.12
	[2.60]		[3.01]	[1.19]	[1.88]		[1.71]	[-0.46]
H	2.93	0.81	2.08	1.02	1.93	0.56	1.10	0.37
	[3.44]		[4.30]	[3.33]	[2.48]		[2.91]	[1.28]
H-L	2.07	1.16	2.01	1.53	1.30	0.71	1.25	0.75
	[4.57]		[4.47]	[3.45]	[2.79]		[2.69]	[1.63]

Fig. 1. Analysis of the predictive power of the composite quality factor based on the least squares construction (self-made)

5 Conclusion

Based on Gordon's dividend growth model, this paper composes company fundamental indicators from five aspects: profitability, growth, corporate governance, potential value and safety, and obtains a data set of company characteristics with 115 indicators. Principal component analysis, Enigma Macbeth regression method, predictive portfolio method, composite principal component analysis and partial least squares method are applied to reduce the dimensionality of the above data set and construct a comprehensive quality index of company fundamentals respectively. Through the univariate portfolio analysis method, it is found that the quality indices constructed based on the partial least squares methodQ^{PLS} has the strongest and relatively stable predictive ability.[10]

Investors confronting too much information are caught in screening difficulties and high-dimensional traps as a result of the big data era. The future could introduce more complex and efficient machine learning algorithms, such as LASSO, Bayesian analysis, etc., to process as much information as possible.

REFERENCES

[1] Fama, E. F., and K. R. French. 1993. Common risk factors in the returns on stocks and bonds, *Journal of Financial Economics* 33, 3-56.

[2] Shleifer, A., and R. Vishny. 1997.The limits of arbitrage. *Journal of Finance* 52, 35–55.

[3] Hou, K., C. Xue, and L. Zhang. 2016. A comparison of new factor models, *NBER Working Paper*.

[4] Fama, E. F., and K. R. French. 2015. A five-factor asset pricing model, *Journal of Financial Economics* 116, 1-22.

[5] Shleifer, A., and R. Vishny. 1997.The limits of arbitrage. Journal of Finance 52, 35–55.

[6] Hou, K., C. Xue, and L. Zhang. 2016. A comparison of new factor models, NBER Working Paper.

[7] Fama, E. F., and K. R. French. 2015. A five-factor asset pricing model, Journal of Financial Economics 116, 1-22.

[8] Z. Peng, "Stocks Analysis and Prediction Using Big Data Analytics," 2019 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS), 2019, pp. 309-312, doi: 10.1109/ICITBS.2019.00081.

[9] Gupta, Shivam, et al. "Achieving superior organizational performance via big data predictive analytics: A dynamic capability view." Industrial Marketing Management 90 (2020): 581-592.

[10] Jin, Xiaolong, et al. "Significance and challenges of big data research." Big data research 2.2 (2015): 59-64.