RMSC4002
Financial Data Analytics with Machine Learning
An Investigation on relationship between Stock Price and Financial Data through Generalized PCR Model

Cheuk Nam CHUNG [*] Ting Tin MA [†] Wei Zhia KUA [‡]

December 3, 2020

Abstract

It is not hard to imagine that the stock price have a strong relation between the financial ratios. In particular, we consider the stock price as a linear combination of those financial ratios. Under the standard setting of multiple regression, one can also complete the task by using the OLS estimator. However, we believe that the price of a certain stock also depends on the financial ratio of the other stocks. Therefore, we use the Generalized Principle Component Regression (GPCR) to understand the intrinsic relationship between the stock price and financial ratios. We then propose some trading strategies and apply Classification Tree to understand the correctness of price. Some potential extension of the proposed procedure related to Time Series Model will also be discussed.

[*] 1155109131@link.cuhk.edu.hk

[†] 1155109956@link.cuhk.edu.hk

[‡] 1155098495@link.cuhk.edu.hk

1 Introduction

The project is organized as follows. In Section 2, the mathematical notation used frequently in this project is introduced. In Section 3, the detailed procedure for the principal component regression and the choice of the basis function would be outlined. In Section 4, it involves discussion on how to handle the practical data set in prior to applying our proposed procedure. In Section 5, we apply the procedure to our data set and provide some financial interpretation. In Section 6, back-test is conducted to evaluate the performance of the proposed model. In Section 7, application involving classification tree and would be discussed. In Section 8 we would attempt to compare the other methodologies with one proposed in this project. Section 9 concludes the whole project with some possible extension of proposed model.

2 Mathematical Notation

The data set for the project is defined as follow, $D = \{X_t, y_t\}_{t=1,\cdots,n}$, where $X_t \in \mathbb{R}^{q\times p}$, $y_t \in \mathbb{R}^q$ and $t = 1,\cdots,n$. The physical meaning of those symbols are as follow.

- X_t stores value of p types of financial ratios for each stock (Total number of stocks $= q$). Explicitly, $[X_t]_{i,j}$ is the value of j-th financial ratio for the i-th stock at time t.
- y_t stores value of stocks at time t, i.e. $[y_t]_j$ is the value of the j-th stock at time t.

Also, denote $|A|$ as the cardinality of the set A. For convention, we would sometime write

$$X_t = \left[\begin{array}{c} X_t^{(1)} \\ \hline X_t^{(2)} \\ \hline \vdots \\ \hline X_t^{(q)} \end{array}\right] = \left[\ F_t^{(1)} | F_t^{(2)} | \cdots | F_t^{(p)}\ \right],$$

i.e. $X_t^{(i)}$ stores value of those financial ratios for the i-th stock at time t, while the RHS term is the alternative partitions of the matrix X_t.

3 Motivation and Model Setting

3.1 Motivation

In classical setting, one might simply consider the non-parametric regression model

$$[y_t]_j = f(X_t^{(j)}) + \epsilon, \quad t = 1, \cdots, n.$$

and estimate f. One can proceed by density estimation or choosing some basis function $\{\widetilde{f_i}\}$ to estimate f by PCR procedure. However, it might not be sensible enough as we believe that $[y_t]_j$ depends on X_t not only through $X_t^{(j)}$. Also, we believe that there is some dependence structure between different factors. The scatter plot (3) and heat map (4) visualize the dependence and motivates us to study the project.

3.2 Model Setting and Proposed Procedure

Therefore, we generalize the model to be

$$y_t = f(X_t) + \epsilon_t,$$

where $f : \mathbb{R}^{q\times p} \to \mathbb{R}^q$ is a matrix function. In general, it is hard to directly estimate f and hence we propose some basis function to estimate f. However, as each of $f_i : \mathbb{R}^{q\times p} \to \mathbb{R}$ is a high dimensional function and hence hard to deal with, we first reduce the dimension of the X_t by PCA and take the first 3 loadings only. The reason of taking the first three loadings only is to retain the financial interpretation (i.e. model the parallel shift, and the curvature). Notice that in the following we kind of using the same notation $f \in \mathbb{R}^{q\times 3} \to \mathbb{R}^q$ and $f_i \in \mathbb{R}^{q\times 3} \to \mathbb{R}$ but they are NOT same as that mentioned before, they are functions applied to the compressed data matrix instead. Denote the resulting compressed data matrix obtained by using the first three loadings as $Z_t = [Z_t^{(1)} | Z_t^{(2)} | Z_t^{(3)}] \in \mathbb{R}^{q\times 3}$. In particular, denote $\alpha^{(i)} = (\alpha_{i1}, \cdots, \alpha_{ip})^T$ as the i-th PC loadings we can evaluate

$$Z_t^{(i)} = \sum_{j=1}^{p} \alpha_{ij} F_t^{(j)}, \quad \text{for } i = 1, 2, 3.$$

Hence the task becomes estimating

$$y_t = f(Z_t) + \epsilon_t = \begin{pmatrix} f_1(Z_t) \\ f_2(Z_t) \\ \vdots \\ f_q(Z_t) \end{pmatrix} + \epsilon_t. \tag{1}$$

We assume that we can write

$$f(Z_t) = g_1(Z_t^{(1)}) + g_2(Z_t^{(2)}) + g_3(Z_t^{(3)}), \tag{2}$$

i.e. a sum of functions of each of Principle Component, where $g_i : \mathbb{R}^q \to \mathbb{R}^q$ for $i = 1, 2, 3$. Base on theory in fourier analysis, we believe that the function f and hence g_1, g_2, g_3 can be well-approximated by sine and cosine functions. Consider

$$\mathscr{S} \coloneqq \left\{ x^k : k \in \mathbb{N} \right\} \quad \text{and} \quad \mathscr{C} \coloneqq \left\{ x^{-k} : k \in \mathbb{N} \right\}$$

Let $\Gamma_1, \cdots, \Gamma_k \subseteq \mathbb{N}^q$ be a sequence of ordered-tuples without repetitions under the constraint that $|\Gamma_i| = q$ and $\Gamma_i \neq \Gamma_j$ for all $i \neq j$. We first choose a sufficiently large number of basis functions to be used, denoted by K. We then define the basis functions $S_1, \cdots, S_K, C_1, \cdots, C_K$ in the following way. For $\Gamma_i = (\gamma_{i1}, \cdots, \gamma_{iq}) \in \mathbb{N}^q$ and $U = (u_1, \cdots, u_q) \in \mathbb{R}^q$,

$$S_i(U) = S(\Gamma_i)(U) \coloneqq \begin{pmatrix} u_1^{\gamma_{i1}} \\ u_2^{\gamma_{i2}} \\ u_3^{\gamma_{i3}} \\ \vdots \\ u_q^{\gamma_{iq}} \end{pmatrix} \quad \text{and} \quad C_i(U) = C(\Gamma_i)(U) \coloneqq \begin{pmatrix} u_1^{-\gamma_{i1}} \\ u_2^{-\gamma_{i2}} \\ u_3^{-\gamma_{i3}} \\ \vdots \\ u_q^{-\gamma_{iq}} \end{pmatrix}$$

We then consider the following assumption

$$g_j(U) = \sum_{i=1}^{K} \Big[a_{ij} S_i(U) + b_{ij} C_i(U) \Big], \quad \text{for } j = 1, 2, 3. \tag{3}$$

Therefore, the task (2) reduces to estimation of $a'_{ij}s$ and $b'_{ij}s$. There are several technical and practical problem to deal with, we then list and provide some possible remedy to them. Also, some further motivations for using such type of approximation.

1. In practice, we cannot set the value of K, i.e. the number of basis function to be arbitrarily large. Instead, we fix a number $N = 50$ first and define $\mathscr{S}_N = \{\sin(2\pi kx) : k = 1, \cdots, N\} \subseteq \mathscr{S}$ and $\mathscr{C}_N = \{\cos(2\pi kx) : k = 1, \cdots, N\} \subseteq \mathscr{C}$, notice that number of permutation of choosing q elements within N elements is given by P_q^N.

2. There are many possible way to choose the basis function. For example, sine, cosine and polynomial functions. However, polynomial function being the basis is One can use sine and cosine functions are suitable for modelling those effects. Meanwhile, by the Taylor Expansion, we can write the polynomial as linear combination of sine and cosine functions and hence somehow applying polynomial basis can also understand the infinitesimal changes.

3. Indeed, it is natural to believe that $[y_t]_j$ depends on Z_t through NOT only $[Z_t]_{j,1}$, $[Z_t]_{j,2}$ and $[Z_t]_{j,3}$, but mainly depends on them. Therefore, one may propose the following extension

$$[y_t]_j = f_j(Z_t) = \lambda A_j([Z_t]_{j,1}, [Z_t]_{j,2}, [Z_t]_{j,3}) + (1 - \lambda) B_j\Big(Z_t \backslash \{[Z_t]_{j,1}, [Z_t]_{j,2}, [Z_t]_{j,3}\}\Big),$$

   where A_j, B_j is also functions to be estimated by using Fourier series with same methods as above. The idea is that as the major information on $[y_t]_j$ comes from $[Z_t]_{j,1}, [Z_t]_{j,2}, [Z_t]_{j,3}$, we can separate the task into estimation of two functions A_j and B_j, where A_j contains information in the main source. The parameter λ is somehow like a penalty factor that add a larger weighting to A and less weighting to B. However, it is easy to see that optimizing $a'_{ij}s$ and $b'_{ij}s$ has automatically attained this task, i.e. we have already accounted for the importance of different information in the proposed model.

4. Simply using Multiple Regression Model by using sine and consine functions as the

basis may also attain the task, i.e. consider total of q models separately

$$[y_t]_j = g_j(X_t^{(j)}), \quad t = 1, 2, \cdots, n$$

for $j = 1, 2, \cdots, q$ and attempted to estimate **different** functions $g_1, \cdots, g_q$. However, it simply uses the information given by its financial ratio but not the others, countering our initiative to study this project. Indeed, it is the reason why we have to consider the permutation setting above for construction of $S_1, \cdots, S_k, C_1, \cdots, C_k$. If we apply the same function in each argument, say

$$S_i(U) \coloneqq \begin{pmatrix} u_1^i \\ u_2^i \\ u_3^i \\ \vdots \\ u_q^i \end{pmatrix} \quad \text{and} \quad C_i(U) \coloneqq \begin{pmatrix} u_1^{-i} \\ u_2^{-i} \\ u_3^{-i} \\ \vdots \\ u_q^{-i} \end{pmatrix}$$

Then it exactly reduces to the case of multiple regression setting as above. Therefore, we consider the permutation of the sine and cosine functions to construct the basis functions. We interpret that optimizing $a_{ij}'s$ and $b_{ij}'s$ under such setting can take the other random effects into account.

We then move back to the procedure. Noticing some drawback of the above model

- Solving $a_{ij}'s$ and $b_{ij}'s$ in (3) is difficult and computation intensive.
- Hard to interpret the effect of different factors given large value of K.

Therefore, we proceed by Principle Component Regression to reduce the data dimension. Also, it is likely that the original $S_1, \cdots, S_K, C_1, \cdots, C_K$ may suffer from the problem of multicollinearity, the estimated coefficient a_{ij}, b_{ij} would be of high variability while aplying

PCA can get rid of this problem. By (2) and (3), we can write

$$\begin{aligned}
y_t =& g_1(Z_t^{(1)}) + g_2(Z_t^{(2)}) + g_3(Z_t^{(3)}) = \sum_{j=1}^{3}\left(\sum_{i=1}^{K}\left[a_{ij}S_i(Z_t^{(j)}) + b_{ij}C_i(Z_t^{(j)})\right]\right)\\
=& \sum_{i=1}^{K}\left[a_{i1}S_i(Z_t^{(1)}) + b_{i1}C_i(Z_t^{(1)})\right] + \left[a_{i2}S_i(Z_t^{(2)}) + b_{i2}C_i(Z_t^{(2)})\right] + \left[a_{i3}S_i(Z_t^{(3)}) + b_{i3}C_i(Z_t^{(3)})\right]\\
=& \sum_{i=1}^{K} a_{i1}S_i(Z_t^{(1)}) + a_{i2}S_i(Z_t^{(2)}) + a_{i3}S_i(Z_t^{(3)}) + b_{i1}C_i(Z_t^{(1)}) + b_{i2}C_i(Z_t^{(2)}) + b_{i3}C_i(Z_t^{(3)})\\
=:& \sum_{k=1}^{6K} d_k g_k(Z_t),
\end{aligned}$$

where

$$\Big(d_k, g_k(Z_t)\Big) = \begin{cases} \Big(a_{k,1}, S_k(Z_t^{(1)})\Big) & , \ k \in [1, K];\\ \Big(a_{k-K,2}, S_{k-K}(Z_t^{(2)})\Big) & , \ k \in [K+1, 2K];\\ \Big(a_{k-2K,3}, S_{k-2K}(Z_t^{(3)})\Big) & , \ k \in [2K+1, 3K];\\ \Big(b_{k-3K,1}, C_{k-3K}(Z_t^{(1)})\Big) & , \ k \in [3K+1, 4K];\\ \Big(b_{k-4K,2}, C_{k-4K}(Z_t^{(2)})\Big) & , \ k \in [4K+1, 5K];\\ \Big(b_{k-5K,3}, C_{k-5K}(Z_t^{(3)})\Big) & , \ k \in [5K+1, 6K]; \end{cases} \tag{4}$$

Notice that even there is n time point, we would **NOT** take all data from $t = 1$ to n into account as the data far long time ago is likely to be useless or even harmful to the analysis. Therefore, we introduce the bandwidth parameter l and focus on understanding y_t through $Z_{t-l+1}, \cdots, Z_t$. In particular we aims to understand y_n in the following discussion (Would be generalized later). We consider the following data matrix:

$$\widetilde{\mathbb{D}} = \Big(g(Z_{n-l+1}) \cdots g(Z_n)\Big), \quad \text{where} \quad g(Z_t) \coloneqq \Big(g_1(Z_t) \ \ g_2(Z_t) \cdots g_{6K}(Z_t)\Big)^T$$

with which the corresponding $6K \times 6K$ sample covariance matrix $\widetilde{S}$ is constructed. Denote the spectral decomposition of $\widetilde{S}$ as $\widetilde{S} \coloneqq HDH^T$ such that $H = (\boldsymbol{h_1}, \boldsymbol{h_2}, \cdots, \boldsymbol{h_{6K}})$, where $\boldsymbol{h_j} = (h_{1j}, h_{2j}, \cdots, h_{6K,j})^T$ is the unit eigenvector associated to the j-th largest eigenvalue λ_j of $\widetilde{S}$, and the corresponding j-th principal component of Z_t, i.e. $P_j(Z_t)$ for $j = 1, 2, \cdots, 6K$

can be computed by $P_j(Z_t) = \sum_{k=1}^{6K} h_{kj} g_k(Z_t)$.

$$y_t = \sum_{j=1}^{M} d_j g_j(Z_t) = \sum_{j=1}^{M} \beta_j P_j(Z_t) + \epsilon_t \tag{5}$$

, where $M << 6K$. $(\beta_1, \cdots, \beta_M)$ can be estimated by OLS estimator for the model in (5). Explicitly, solve

$$\hat{\beta} = \left[(P^{(M)})^T P^{(M)} \right]^{-1} \left(P^{(M)} \right)^T Y,$$

where $P^{(M)} \in \mathbb{R}^{nq \times M}$ is a large matrix with $[P^{(M)}][n(t-1)+1 : nt, 1 : M] = [P_1(Z_t) | \cdots | P_M(Z_t)]$. Similarly, $Y = [y_1^T | \cdots | y_n^T]^T$.

3.3 Brief Summary of Procedure

```
Algorithm 1: Generalized Principle Component Regression (GPCR)
Input:
(i) $D = \{X_t, y_t\}_{t=1,\cdots,n}$, the data set required.
(ii) Value of $l, N, M, K$. (Determined by user; Affects computational cost)
(iii) $X_1, \ldots, X_n \in \mathbb{R}^{q \times p} - n$ data matrix, the $(i,j)$-th entry of $X_t$ refers to the value of $j$-th
    financial ratio for the $i$-th stock at time $t$.
begin
    (1) For each $X_t$ $(t = 1, 2, \cdots, n)$, apply PCA to obtain the first three loadings and the
        associated PCs $Z_t^{(1)}, Z_t^{(2)}, Z_t^{(3)}$ and the resulting compressed data matrix is denoted
        by $Z_t \coloneqq [Z_t^{(1)} | Z_t^{(2)} | Z_t^{(3)}]$.
    (2) Decide the value of $N$, meaning that we would use some of $x^j$ and $x^{-j}$ (if all
        entry are non-zero) for $j = 1, 2, \cdots, N$ as the univariate basis functions. Fix $K << P_q^N$
        as the number of the multivariate basis function to be used, where each of them are
        constructed by the univariate one.
    (3) Consider $\widetilde{\mathbb{D}} = \big( g(Z_1) \cdots g(Z_n) \big)$ as defined in (4), evaluate the sample covariance
        matrix $\widetilde{S}$ of $\widetilde{\mathbb{D}}$.
    (4) Apply spectral decomposition on $\widetilde{\mathbb{D}}$ to obtain $\widetilde{\mathbb{D}} = HDH^T$, where
        $H = (\boldsymbol{h_1}, \boldsymbol{h_2}, \cdots, \boldsymbol{h_{6K}})$ and $\boldsymbol{h_j} = (h_{1j}, h_{2j}, \cdots, h_{6K,j})^T$ is the unit eigenvector
        associated to the $j$-th largest eigenvalue $\lambda_j$ of $\widetilde{S}$.
    (5) Set $P_j(Z_t) = \sum_{k=1}^{6K} h_{kj} g_k(Z_t)$ for $j = 1, \cdots, 6K$.
    (6) Fix $M << 6K$. Consider the model $y_t = \sum_{j=1}^{M} \beta_j P_j(Z_t)$, obtain coefficients $\beta_j' s$ by
        the OLS estimators.
    (7) Set $d_i^{(M)} = \sum_{j=1}^{M} h_{ij} \beta_k$.
return: The coefficients $d_1^{(M)}, \cdots, d_{6K}^{(M)}, \beta_1, \cdots, \beta_M$ for the model in (5).
```

We make the following remark to the above procedure.

- ***Important :*** If there exist zero entry in the data matrix, we would simply use S_i only as C_i which involves reciprocal of zero is NOT well-defined.
- N is set to be 5 as usually the financial data can be explained by degree 5 polynomial.
- Commonly, we take $M << 6K$ and estimate $\hat{y}_t \coloneqq \sum_{j=1}^{M} \beta_j P_j(Z_t)$ as most of the variability of the data set is retained in the first M PCs.
- This procedure generalize the standard PCR in the following sense. (1) The basis function maps a data matrix X_t to $\mathbb{R}^q$, while the standard functions can only handle $\mathbb{R}^p \mapsto \mathbb{R}^q$. (2) The procedure is indeed a two stage-PCA. Applying PCA for X_t each time point followed by applying PCA on the fitted model in (5).

4 Data Cleansing Processing Procedure

Daily stock data from 2019-01-01 to 2020-11-27 are collected from the Bloomberg terminal. We first obtain stock price of 1128 stocks listed in Hong Kong Stock Exchange during the desired period, and eventually narrow down the number of stocks to 321 due to missing data.

There are two major factors of the missing data problem — unavailability of data from Bloomberg and the error triggered by the *BDH()* function, the Bloomberg add-in excel command which returns data values. Sufficient time is needed for the command to export data from Bloomberg. If a large number of commands are called at the same time, some of them will not return a single value. The procedure of handling missing data is as follows.

Two approaches are taken to handle the missing data. Since 471 days data are obtained, for a single datum loss, average will be taken from the previous datum and the next datum. For data loss in consecutive days, linear interpolation will be used to infer the missing data. If the data loss includes the first date or the last date of the financial data of the stock, the whole stock will be deleted and will not be under our investigation as it is hard to infer the behaviour of the data trend.

Also, stocks with first day price less than HK$1 are removed. Since some of the penny stocks price movements are nearly constant but with abnormal fluctuation over a short period of time, such correlations between the price and its financial data cannot be captured and hard to contribute to the general picture of how the financial data determine the stock price.

After that, we first calculate the number of stocks with full data available of each ratio, then use the stocks in the least number stocks available data. Finally, we check if other financial data of those stocks are complete sets with no N/A values. If so, that particular stocks will be in use for analysis. There are in total 185 stocks sorted out from the above procedures.

Besides, data types like market capitalization and trading volume are significantly larger than the corresponding stock price which may hinder the principal component analysis by being the dominating factors, taking logarithmic value can avoid the stress on those factors.

Including the stock price dataset, there are 10 csv files in total. To facilitate data processing, we rename the stock price dataset file with alphabet Z be the first letter so that it can be listed as the last item by the *R* command *list.flies()*.

	Financial Data Type
1	Average Bid-ask Spread
2	Closing Price 1 day before
3	Earning Yield
4	Market Capitalization
5	Overridable Adjusted Beta
6	Overridable Alpha
7	Price to Book Ratio
8	Price to Sales Ratio
9	Trading Volume

Figure 1: Financial Data chosen for Principle Component Regression

5 Interpretation of Implemented Result

After performing the principal component regression along a moving time frame with length 30, we obtain the estimated stock prices ($\hat{y}_t$) of each chosen stock in each time point which are the intrinsic prices we believe that the stock has.

5.1 Pricing Indicator

It serves as the buying signal of a stock. If the estimated stock price (intrinsic value) is lower than the true stock price ($\hat{y}_t < y_t$), we regard the stock as overpriced and one should sell or hold the stock in short position. On the contrary, if the estimated stock price (intrinsic value) is higher than the true stock price ($\hat{y}_t > y_t$), we regard the stock as underpriced and one should buy or hold the stock in long position. Graphically speaking, from (2), the stock data points below the reference line are regarded as overpriced, while those being above the reference line are regarded as underpriced. The graphical interpretation is similar to the graph plotting expected return against stock beta with the security market line, which is the graphical representation of the classic capital asset pricing model.

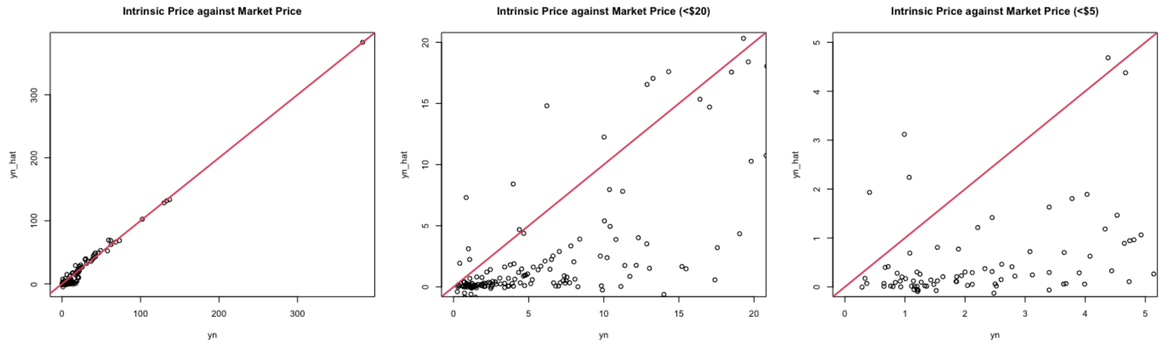

Figure 2: Scatter plots of intrinsic price versus market price (All, <$20, <$5)

5.2 Error Analysis

From (2), there are large discrepancy between the predicted intrinsic stock prices and their market prices when the price is small. A possibLe explanation is because one of the factors is stock price 1-day before. When the stock price is large, its stock price 1-day before is reasonably large as well and they are significantly larger than any else of the factors. Therefore, the high correlation between them are revealed in the estimated function. However, when the stock price is small, its stock price 1-day before is small as well and they are not significantly different from other factors. There is no dominance of any factors. One may suspect if the stock priced predicted is biased due to the dominating factor. However, as we perform this analysis for detecting the overpricing or underpricing of stocks. (Refer to (6)) The magnitude does not matter much and the model can still be utilized.

5.3 Classification Tree Analysis

As shown in the cross tabulation table below, the misclassification rate of this model is around 11.4%. From the classification tree (5), we observe that the stock is likely to be overpriced if the company has small market value and low earnings yield. Small cap companies with high P/E ratio are often referred to as growth stocks since the earnings growth is expected to be high. This is consistent with the fact that company with high growth potential will attract more investors and hence it is likely to be overpriced. On the other hand, we notice from the plot that if the bid-ask spread is high, the stock is likely to be underpriced. Recall that bid-ask spread is designed to cover the market risks and costs to yield a reasonable profit for the market maker. This include the costs of processing orders and trading. Besides, illiquid stocks and riskier stocks tend to have larger bid-ask spread. Even though bid-ask spread is seldom used to determine a stock's intrinsic value, our investigation reveals that stocks with high bid-ask spread are more likely to be underpriced. This suggests that market maker can influence stock valuations.

pr	0	1
Underpriced	18	2
Overpriced	19	146

6 Back-Test of the Proposed Procedure

We do backtesting on the proposed trading strategy algorithm on a stream of historical financial ratios to generate a set of trading signals, each representing a profit and loss. We recognize that there exists many biases that may affect the backtesting performance. To mitigate data-snooping bias, we consider using large quantity of data sets in the training set to avoid overfitting.

We find out that the success rate of predicting a stock price trend is 54.44%. The probability is not so high but it is greater than 50%. Indeed, if we apply this as an indicator (refer to (5.1)) of buying or selling a stock in each time period (7 trading days), we can capture a profit of 9.68%. While a passive buy and hold strategy of buying the indicated stocks in the first time point and selling the stock in the last time point captures a loss of 2.85%, while an average portfolio with buy and hold strategy obatin a 2.08% profit. This shows that the indicator is effective in generating a larger return of the portfolio. However, as the frequency of trading stocks increase substantially, one may also be aware of the transaction costs that may lower the return obtained.

7 Potential Applications of the Proposed Procedure

7.1 Detection of over/under-pricing

We can regard the estimated $\hat{y}_t$ by GPCR as a guidance to price the stocks at time t. If $[y_t]_j > [\hat{y}_t]_j$, we claim the j-th stock to be over-priced. If $[y_t]_j < [\hat{y}_t]_j$, we claim it to be under-priced. While $[y_t]_j \neq [\hat{y}_t]_j$, the price is sensible. However, the actual price is almost surely NOT equal the predicted price, so commonly $[y_t]_j \neq [\hat{y}_t]_j$. Therefore, we can consider the following quantity. Commonly, investors claimed that a stock with low P/E ratio is likely to be under-priced because P/E ratio indicates the willingness of purchase the company's stock based on the reported earnings.
Some other financial ratios might also affect the appropriateness of pricing, for example, price-to-book ratio and earning yield. Therefore, it motivates us to consider the following quantity

$$Q_j \coloneqq 1([y_n]_j - [\hat{y}_n]_j), \quad \text{for } j = 1, 2, \cdots, q,$$

i.e. whether the q different stocks are over/under priced by using our estimated price as the reference price $\mathbb{Q} \coloneqq (Q_1 \cdots Q_q)^T \in \mathbb{R}^q$. Our interested data set is hence given by (Q, X_n). In here, we partition the matrix X_n in another way by $X_n = [F_1|F_2|\cdots|F_p]$. We then consider the classification tree of Q with respect to the factors $F_1, \cdots, F_p$. Given our model is true, we want to understand whether a stock is over- or under-priced. Refer the result for Classification Tree to (5) and it states the following classification rules: The stock is likely to be over-priced if either

1. "Logarithm of the Market value ≥ 11.49" and "Average Bid Ask Spread ≥ 0.05455".

2. "Logarithm of the Market value ≤ 11.49" and "Earning Yield < 26.14".

Otherwise, the stock is likely to under-priced.

8 Comparison with Existing Methodologies

8.1 Multiple Linear Regression

As claimed in the beginning of the project, we consider that $[y_t]_j$ depends on X_t NOT only through $X_t^{(j)}$. We attempted to verify this claim. Consider the following q different regression models

$$[y_t]_j = M_j(X_t^{(j)}) + \epsilon_t = \sum_{i=1}^{p} \alpha_i [X_t^{(j)}]_i + \epsilon_t \quad \text{for } t = 1, \cdots, n. \tag{6}$$

In the above setting, $M_i \neq M_j$ in general for $i \neq j$. We try to compare the value obtained by those q different model. Denote the estimated values under the setting in (6) as $\widetilde{y_t}$ and that obtained by our procedure as $\widehat{y_t}$. It is natural to see that applying our procedure can estimate the stock price better by taking more information into account.

8.2 AR models in each predictors

Indeed, one might consider each of the financial ratio itself forms a time series. For example, denote $\{ES_t^{(j)}\}_{t=1,\cdots,n}$ as the earning per share ratio for the j-th stock. We then apply AR(p) model to predict $ES_{t+1}^{(j)}$. Similarly, we can apply it to all financial ratio to get an estimated value of Z_{n+1}, denoted by $\hat{Z}_{n+1}$. We then estimate $\hat{y}_{n+1} = \hat{f}(\hat{Z}_{n+1})$, where $\hat{f}$ is obtained by our proposed procedure.

9 Conclusions and Discussions

9.1 Summary and Limitations

To summarize, we have proposed the Generalized PCR to handle with higher dimension data set with some possible choice of basis functions. One can apply the model to NOT only financial purpose but many other similar cases in the sense that we believe the value of certain response depends on value of predictors of remaining responses. However, there is some potential drawbacks. First, the computational cost is heavy as it involves PCA for each time points followed by some optimization procedure. Secondly, the stock price may depends heavily on factors not included in proposed model. There is missing data for some financial ratios, the data cleansing procedure also affects the fitted model.

9.2 Possible Extensions and Discussions

9.3 Generalization of choice of g_i

Indeed, recall from equation (4), we did not directly use the value of accounting ratio of other stocks to estimate our stock, but implicitly embedded it into the regression model as we believe that the dependence can be automatically captured as long as K is large enough. Therefore, a natural extension is to directly use value of accounting ratio of other stocks into account but it adds hevay computation burden.

9.3.1 Mixture of Polynomial and Fourier Basis

We consider using the Fourier Basis to construct the Multivariate Basis Function because we believe sine and cosine functions can model infinitesimal changes. However, we can only use finitely many sine and cosine functions for form the basis, we then consider introducing also the polynomial basis into the model, i.e. $y_t = g(Z_t) + f(Z_t) + \epsilon_t =: Pf(Z_t) + \epsilon_t$, where g can be estimated by mimicking Algorithm (1) and it is regarded as an two-step optimization problem as

$$\underset{Pf \in \mathscr{P}}{\arg\min} \sum_{t=1}^{N} (y_t - Pf(Z_t))^2 = \underset{g \in \mathscr{G}}{\arg\min} \left(\underset{f \in \mathscr{F}}{\arg\min} \sum_{t=1}^{N} \Big(y_t - g(Z_t) + f(Z_t) \Big)^2 \right),$$

where $\mathscr{G}$ and $\mathscr{F}$ are the function space induced by algorithm (2) and the class of functions of polynomial basis, while P is that jointly induced by g and f. The motivation is that smooth functions can be approximate by Fourier series, so introducing some polynomial function somehow also helps capture the effects better. For example, one may consider replacing S_i and C_i (The Multivariate Polynomial Basis function) by

$$S_i(U) = S(\Gamma_i)(U) \coloneqq \begin{pmatrix} \sin(2\pi\gamma_{i1}u_1) \\ \vdots \\ \sin(2\pi\gamma_{iq}u_q) \end{pmatrix} \quad \text{and} \quad C_i(U) = C(\Gamma_i)(U) \coloneqq \begin{pmatrix} \cos(2\pi\gamma_{i1}u_1) \\ \vdots \\ \cos(2\pi\gamma_{iq}u_q) \end{pmatrix}$$

9.3.2 Time-Series Model

Indeed, it is natural that y_t depends heavily on $y_{t-1}, \cdots, y_{t-p}$, i.e. an AR(p) model and hence we can consider the modified AR(p) model by taking the financial ratios into account. Explicitly, we write

$$y_t = \alpha_1 y_{t-1} + \cdots \alpha_p y_{t-p} + f(Z_t) + \epsilon_t, \tag{7}$$

where f can be estimated using our proposed procedure together with the `optim` function in `R`. The procedure is outlined in the following:

Algorithm 2: Autoregressive–Generalized Principle Component Regression

```
Input:
(i) Input in Algorithm (1)
(ii) Initialization of the coefficients: $\alpha_1^{(0)}, \alpha_2^{(0)}, \cdots, \alpha_p^{(0)}$.
(iii) The self-written function GPCR=function($(X_t)_{t=1,\cdots,n}$, $(y_t)_{t=1,\cdots,n}$) to implement
  Algorithm (1) and returning the function $\hat{f}$ that fits (1)
begin
    (1) Consider $\breve{y}_t = y_t - \alpha_1^{(0)} y_{t-1} - \cdots - \alpha_p^{(0)} y_{t-p}$ for $t = \boldsymbol{p+1}, \cdots, n$.
    (2) Implement to proposed procedure to $(\breve{y}_t)_{t=p+1,\cdots,n}$ instead, i.e. apply
    $\hat{f}$ =GPCR( $(X_t)_{t=p+1,\cdots,n}, (\breve{y}_t)_{t=p+1,\cdots,n}$ ) to obtain a the desired estimator $\hat{f}$ of $f$.
    (3) As for a fixed choice of basis functions and given the data set, the resulting
      estimating function $\hat{f}$ depends only on $\alpha_1^{(0)}, \alpha_2^{(0)}, \cdots, \alpha_p^{(0)}$. Apply optim to minimize
      the SSE, i.e. $\sum_{t=p+1}^{n} \hat{\epsilon}_t^{\,2} = \sum_{t=p+1}^{n} \left(y_t - \alpha_1 y_{t-1} - \alpha_2 y_{t-2} - \cdots - \alpha_p y_{t-p} - \hat{f}(Z_t)\right)^2$ with
      respect to $\alpha_1, \cdots, \alpha_p$ using the initial guess as $\alpha_1^{(0)}, \alpha_2^{(0)}, \cdots, \alpha_p^{(0)}$.
return: the coefficients $\alpha_1, \cdots, \alpha_p$ together with estimated function $\hat{f}$.
```

9.4 Alternative Trading Strategy

Similar to price momentum strategy, we can buy the best performing stocks and sell the worst performing stocks in the portfolio, where "performance" here depends on the selection criterion based on a linear combination of values. Value can be defined as any financial ratio that is believed to have strong relation with the stock price. Using the proposed model, we get the estimated stock price $[\hat{y}_t]_j$, which we can use to check for the deviations of returns between the estimated price and the actual price. Consider

$$\epsilon_j(t) = [R_t]_j - [\hat{R}_t]_j), \qquad \text{for} j = 1, 2, \ldots, q.$$

We can then use these residuals $\epsilon_j(t)$ to compute the risk-adjusted residual returns $\tilde{R}_j^{\text{risk.adj.}}$ given as below (we consider t to be the time measured in the units of 1 month, since the holding period is typically 1 month.

$$\begin{aligned}
\epsilon_j^{\text{mean}} &= \frac{1}{T}\sum_{\text{t=1}}^{T} \epsilon_j(t) \\
\tilde{R}_j^{\text{risk.adj.}} &= \frac{\epsilon_j^{\text{mean}}}{\tilde{\sigma_j}} \\
\tilde{\sigma} j^2 &= \frac{1}{T-1}\sum_{\text{t=1}}^{T} (\epsilon_j(t) - \epsilon_j^{\text{mean}})^2
\end{aligned}$$

We can construct a dollar-neutral portfolio with $\sum_j^q |w_j| = 1$ by buying stocks in the top decile of $\tilde{R}_j^{\text{risk.adj.}}$ and selling stocks in the bottom decile, where $w_j > 0$ for long stocks and $w_j < 0$ for short stocks. We may consider uniform weights or non-uniform weights given by $w_j \propto \frac{1}{\sigma_j}$ or $w_j \propto \frac{1}{\sigma_j^2}$, etc.

SUPPLEMENTARY MATERIALS

The following is the visualization of the p different financial quantities of total of q different stocks at day n.

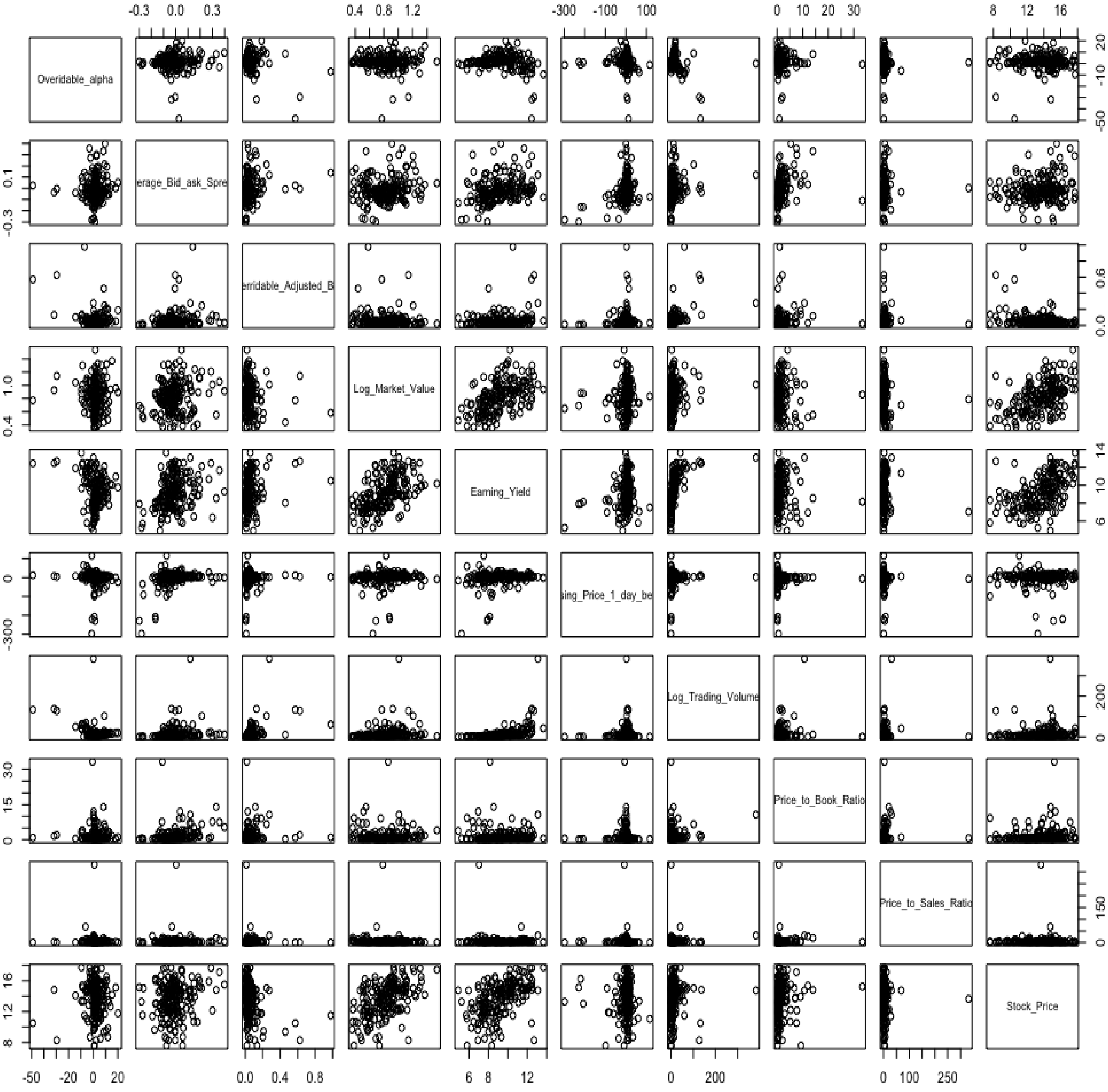

Figure 3: Scatter Matrix.

SUPPLEMENTARY MATERIALS

The attached is the heat map for the data matrix, which indicates the existence of dependence structure in a clearer manner than (3).

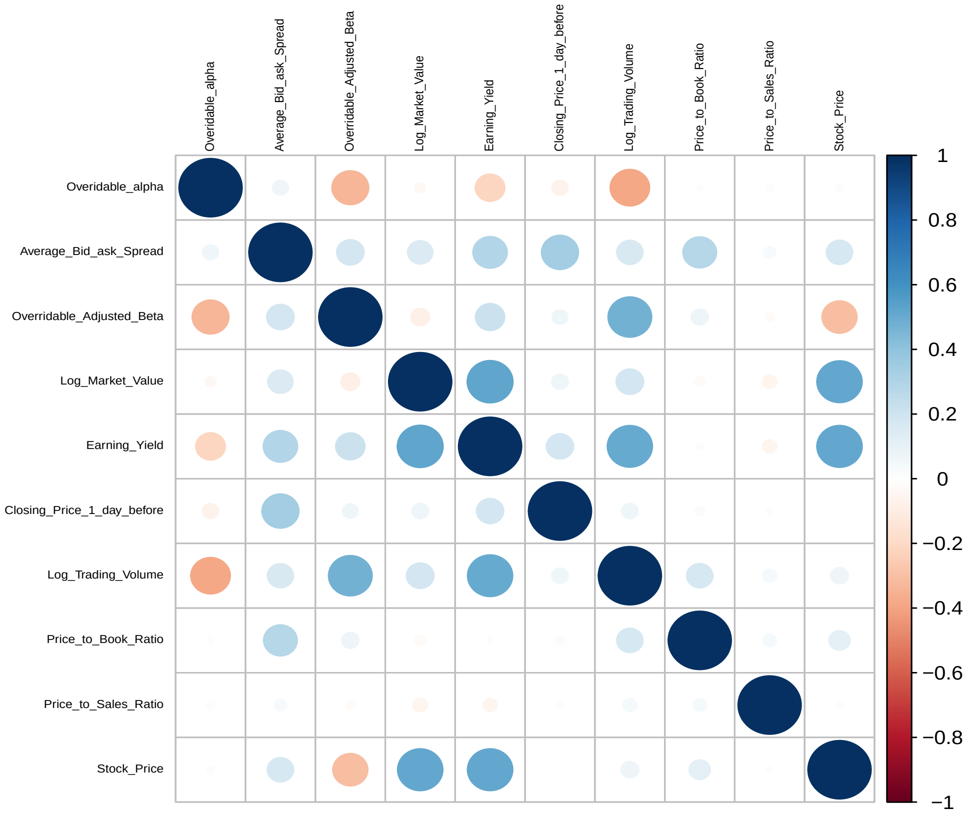

Figure 4: Heat Map of the stock price and 9 factors

SUPPLEMENTARY MATERIALS

The following refer to the Classification Tree of the the over-pricing indicator with respect to other factors. In the following case, we take $N = 7$.

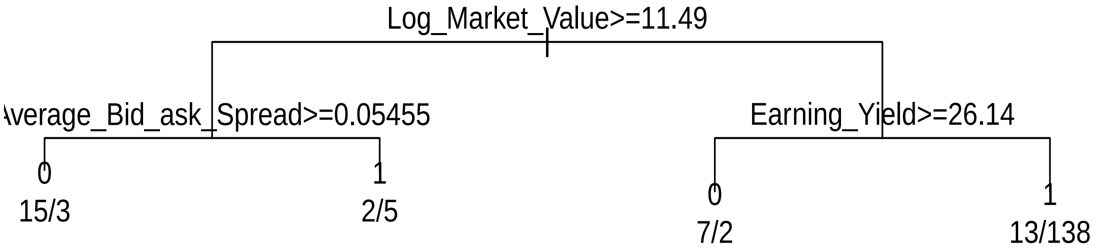

Figure 5: Classification Tree