第51卷第5期 中 国 科 学 技 术 大 学 学 报 Vol. 51, No. 5
2021年5月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA May 2021

Received: 2021-04-07; Revised: 2021-04-13 doi:10.52396/JUST-2021-0102

# Impact of COVID-19 pandemic on stock market via sparse principal component analysis

LI Ming, WEN Canhong*

Department of Statistics and Finance, School of Management, University of Science and Technology of China, Hefei 230026, China

* Corresponding author. E-mail: wench@ustc.edu.cn

**Abstract**: The COVID-19 pandemic has caused severe public health and economic consequences around the world. It is of great importance to evaluate the impact of the COVID-19 pandemic on the economy, especially the stock market. To this end, we proposed to use several state-of-art sparse principal component analysis (PCA) methods for the stock data of the CSI 300 index from February 1, 2019 to February 1, 2021. To show the influence of the outbreak of the COVID-19 pandemic, we divide this period into two periods, i.e., before and after January 1, 2020. Based on this division, we attempted to extract the principal components and construct portfolio accordingly. The results show that the proportion of principal components representing the market declined after the outbreak. For the constitution in the first two principal components, the important stock sets are substantially different after the outbreak. The stocks from the health care sector start to play an important role in the portfolio of the CSI 300 index after the outbreak. Compared with the CSI 300 index, the first two principal components from the sparse PCA methods can obtain higher returns with a much smaller set of stocks in the portfolio. In conclusion, the outbreak of the COVID-19 pandemic led to changes in both proportion and constitution of the principal component of the stocks in the CSI 300 index.

**Keywords**: COVID-19 pandemic ; sparse PCA; stock index

**CLC number**: F830.91 **Document code**: A

## 1 Introduction

### 1.1 Background and data description

The COVID-19 pandemic from the end of 2019 not only poses a serious threat to human life and health, but also causes major losses to the global economy. In order to control the pandemic, quarantine measures have been gradually adopted around the world, some economies have been temporarily shut down, and financial markets have also fallen into a state of continuous decline. As known, the stock market is a barometer of the economy. The stock price not only fluctuates with changes in the economic cycle, but also indicates the economic development situation, and its fluctuations may lead to a recession in the real economy. For example, after the outbreak of the COVID-19 pandemic, on the first day of the opening of the Shanghai and Shenzhen markets, 3188 stocks in the two markets fell by their limit.

Figure 1 shows the CSI 300 index, an index compiled from the most representative 300 stocks in Shanghai and Shenzhen A-shares with large scale and good liquidity. It can be seen that it dropped down 7.88% on February 3, 2020, and experienced several major fluctuations during the COVID-19 pandemic period.

It is of great interest to to evaluate the impact of the COVID-19 pandemic on the Chinese stock market, especially the CSI 300 index, one of the most important benchmarks in the A-share market. However, the CSI 300 index contains 300 stocks from various fields, and the correlation between them is complicated, so it is difficult to analyze them. Thus it is essential to perform some dimension reduction to the 300 stocks and find out the difference before and after the outbreak of the COVID-19 pandemic.

### 1.2 Literature review

Principal component analysis (PCA) is one of the most popular technique to reduce the dimension, and has been well studied in the fields of statistics and finance, such as portfolio management[1]. By converting stocks into a new set of uncorrelated principal components that represent uncorrelated risk sources, PCA can reduce the complexity of stock portfolios. PCA allows us to

**Citation**: LI Ming, WEN Canhong. Impact of COVID-19 pandemic on stock market via sparse principal component analysis. J. Univ. Sci. Tech. China, 2021, 51(5): 404-418,430.

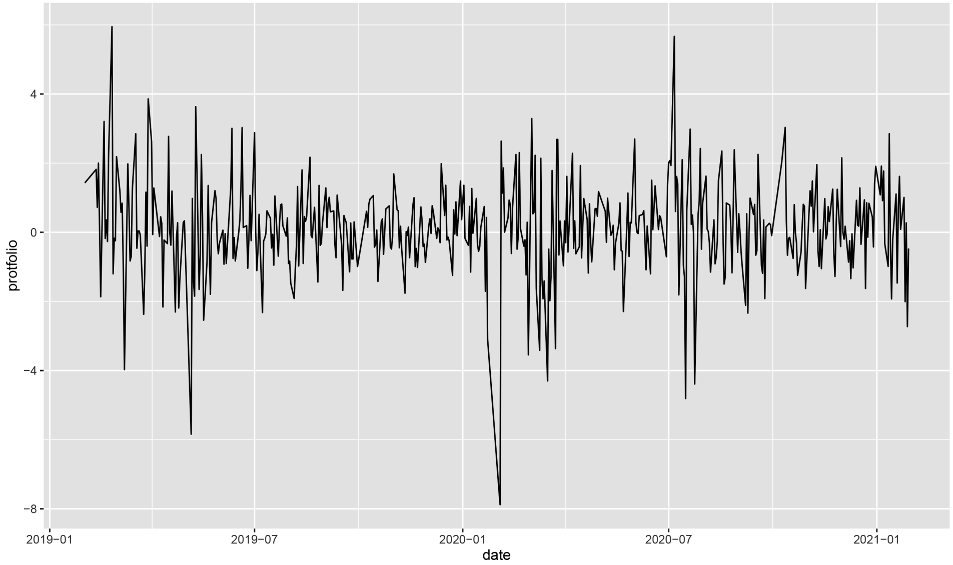


**Figure 1.** The CSI 300 index portfolio from January 1, 2019 to February 1, 2021.

determine the stocks that can be used as a representative of the entire data set, thereby finding the number of stocks that are sufficient to diversify the portfolio.

Due to the large number of stocks, the principal components (PCs) derived from PCA consist of all stocks and it is hard to explain the effect of each stock. To address this issue, sparse version of PCA has been proposed by controlling the number of the variables in the PCs.

The first class of approaches are based on ad-hoc methods by post-processing the PCs obtained from the standard PCA. For example, Cadima and Jolliffe[2] proposed a simple thresholding approach by artificially setting the loadings with absolute values smaller than a threshold to zero. Vines[3] considered simple principal components by restricting the loadings to take values from a small set of allowable integers such as 0, 1, and −1. Although these methods are simple to operate, the sparse PCs obtained usually have large errors. In 2013, Ma[4] proposed an iterative thresholding sparse PCA algorithm based on the QR decomposition. It overcomes the drawbacks of simple thresholding by iterative update, and thus has better performance in both theory and numerical experiment. As the authors stated, however, the convergence of the algorithm is not guaranteed.

In recent years, more involved approaches have been presented. These methods usually impose regularization on different PCA formulations. A classic perspective is that PCA finds a set of directions (technically, a linear subspace) that maximizes the variance of the data once it is projected into that space. Jolliffe et al.[5] developed the Simplified Component Technique-LASSO (SCoTLASS) algorithm for finding sparse orthogonal loading vectors by sequentially maximizing the approximate variance explained by each PC under the $l_1$-norm penalty on loading vectors. d'Aspremont et al.[6] proposed a direct sparse PCA (DSPCA) method, to obtain sparse PCs by solving a sequence of semi-definite program relaxations. In 2008, Journée et al.[7] formulated the sparse PCA problem as a nonconcave maximization problem with $l_0$-norm or $l_1$-norm sparsity-inducing penalties. They showed that the problem can be reduced into maximization of a convex function on a compact set, and developed a computationally efficient gradient method for finding a stationary point. Inspired by the greedy method for solving the combinatorial problem[8], d'Aspremont et al.[9] proposed a greedy heuristic algorithm to solve a new sparse PCA semi-definite programming problem. In 2013, Croux et al.[10] proposed to use a grid search algorithm to derive a sparse version of PCs and achieved desirable results. The eigen decomposition formulation of PCA also relates PCA to the singular value decomposition (SVD) of data matrix. Shen and Huang[11] used SVD to calculate the low-rank matrix approximation of the data matrix under various sparsity-inducing penalties, and applied it to the sparse PCA problems (sPCA-rSVD). This method uses least squares linear regression and simple threshold rules, so it is relatively easy to implement.

Alternatively, the PCs can be interpreted in geometric as the closest linear manifold approximation to the observed data. Based on this, Zou et al.[12] formulated the sparse PCA problem as a regression-type optimization problem and imposed constraint on the coefficients via a combination of $l_1$- and $l_2$-norm penalties. PCA can be also reformulated as a maximum likelihood solution to a latent variable model, called probabilistic PCA[13]. For example, Sigg and Buhmann[14] modified the expectation maximization approach of probabilistic PCA to encourage sparsity or non-negativity to the loading factors.

In this article, we aim to investigate the structural changes of the index stocks before and after the outbreak of the COVID-19 pandemic. To this end, we collect the constituent stock data in the CSI 300 index from the Wind database ( http://www. wind. com. cn/), and propose to apply sparse PCA techniques as well as the classic PCA method to the index stocks data. We also examine the performance of different sparse PCA methods in extensive simulated data.

The structure of the remaining contents is as follows. Section 2 introduces several commonly used sparse PCA methods. Section 3 shows the analysis on CSI 300 constituent stock data via PCA and sparse PCA technique. In Section 4, numerical experiments are carried out on several commonly used sparse PCA algorithms. We conclude with a short discussion in Section 5.

## 2 Algorithms for the sparse PCA problem

In this section, we review four popular sparse PCA methods, which include the variance maximization ( VM )[10], the reconstruction error minimization ( REM )[12], the singular value decomposition (SVD)[11] and the probabilistic model (PM)[14]. Table 1 gives a brief summary on the four methods, and we will discuss them in details in the afterward sections.

Before discussing the methods, we define some notations. Denote the data we have collected as $\{x_i, i = 1, \cdots, n\}$ and its corresponding matrix form as $X \in \mathrm{R}^{n\times p}$. Without loss of generality, assume that the matrix $X$ has been centered to have zero mean. Let $\mathrm{rank}(X) = r$. Let $k$ be the maximum number of (sparse) principal component (PC) to be calculated. Let $v_j$ be the first $j$-th (sparse) loading vector, and $V = (v_1, \cdots, v_k)$ be the loading matrix. Then the corresponding PC is $Xv_j, j = 1, \cdots, k$. Let $I_k$ denote the $k\times k$ identity matrix. For any matrix $M = (M_{ij})$, define the Frobenius norm of $M$ as $\| M \|_F = \sum_i \sum_j M_{ij}$.

### 2.1 Variance maximization (VM)

The variance maximization approach was proposed by Croux et al.[10]. The main idea is to preserve as many changes in the original sample as possible when projecting data points into a low-dimensional space. Note that maximizing the variance of the first PC $Xv_1$ can be equivalently expressed as

$$\max_{v_1} v_1^{\mathrm{T}} X^{\mathrm{T}} X v_1 \quad \text{s. t.} \quad v_1^{\mathrm{T}} v_1 = 1 \tag{1}$$

Based on this formulation, Croux et al.[10] introduced a sparse PCA framework by imposing a $l_1$ penalty function into the objective function of (1), that is

$$\max_{v_1} v_1^{\mathrm{T}} X^{\mathrm{T}} X v_1 - \lambda_1 \| v_1 \|_1 \quad \text{s. t.} \quad v_1^{\mathrm{T}} v_1 = 1 \tag{2}$$

where $\lambda_1$ ($\geqslant 0$) is a regularization parameter that controls the amount of shrinkage on the first PC. It can be seen that the larger the value of $\lambda_1$, the greater the amount of shrinkage (i. e. the greater the amount of zero estimates). Then based on the top $(j-1)$-th PCs, the $j$-th sparse PCs can be solved via the following optimization problem:

**Table 1.** Summary of sparse PCAs: VM, REM, SVD, and PM.

| Method | Formulation | Available package in R | Reference |
|---|---|---|---|
| VM | $\max_{v_j} v_j^{\mathrm{T}} X^{\mathrm{T}} X v_j$<br>$\text{s. t.} \quad v_j^{\mathrm{T}} v_j = 1,$<br>$v_1^{\mathrm{T}} v_i = 0,\ i = 1, \cdots, j-1,$<br>$\| v_1 \| < t$ | pcaPP | Croux et al.[10] |
| REM | $\min_{A,B} \sum_{i=1}^{n} \| x_i - AB^{\mathrm{T}} x_i \|^2$<br>$\text{s. t.} \quad A^{\mathrm{T}} A = I_k,$<br>$\sum_{j=1}^{k} \| B_j \| < s$ | elasticnet | Zou et al.[12] |
| SVD | $\min_{U,D,V} \| X - UDV^{\mathrm{T}} \|_F^2 + \sum_{j=1}^{k} \lambda_j \| V_j \|_1$<br>$\text{s. t.} \quad U^{\mathrm{T}} U = I_k,$<br>$V^{\mathrm{T}} V = I_k$ | PMD | Witten et al.[15],Shen and Huang[11] |
| PM | $X = WZ + \varepsilon$<br>$f(Z_i) \sim N(0, I_k)$<br>$f(\varepsilon) \sim N(0, \sigma^2 I_p)$ | nsprcomp | Sigg and Buhmann[14] |

$$\begin{aligned} \max_{v}\ & v^{\mathrm{T}}X^{\mathrm{T}}Xv-\lambda_j \parallel v \parallel_1 \\ \text{s. t.}\ & v^{\mathrm{T}}v=1, \\ & v^{\mathrm{T}}v_1=0, \\ & \cdots \\ & v^{\mathrm{T}}v_{j-1}=0 \end{aligned} \tag{3}$$

where $\lambda_j$ ($\geqslant 0$) is a regularization parameter that controls the amount of shrinkage on the first $j$-th PC.

Since the principal components are required to be orthogonal, $j-1$ constraints are added for solving the $j$-th PC, which makes the problem (3) difficult to solve. Croux et al.[10] proposed to use a grid search algorithm to obtain the PCs, and showed that it is very fast and has a high accuracy in the high-dimensional setting. The basic idea of the algorithm is to simplify the problem into a series of optimizations in a two-dimensional plane under the constraint of unit norm. This boils down to a series of maximization of functions on the unit circle. This is simply a univariate maximization problem, which can be solved by a grid search.

Before presenting the algorithm, we first combine the problems (2) - (3) into a single optimization problem. Denote the top ($j-1$)-th estimated loading matrix as $\widehat{V}_{j-1}=(\widehat{v}_1,\cdots,\widehat{v}_{j-1}), 1\leqslant j\leqslant k-1$, where $\widehat{v}_1$ is the $i$-th estimated loading vector. For $j>1$, let $\widehat{V}_{j-1}^{\perp}$ be a matrix whose columns are a set of orthogonal bases of subspaces orthogonal to the space generated by the column of $\widehat{V}_{j-1}$. In order to maintain symbol consistency, let $\widehat{V}_0^{\perp}$ be equal to the identity matrix. Define $X^{(j-1)} = X\widehat{V}_{j-1}^{\perp}$, obviously $X^{(j-1)}$ belongs to the low-dimensional space $\mathbb{R}^{p-j-1}$. Then, the optimization problems (2) - (3) are equivalent to

$$\begin{aligned} \max_{v} f(v) &= v^{\mathrm{T}}X^{(j-1)\mathrm{T}}X^{(j-1)}v - \lambda_j \parallel \widehat{V}_{j-1}^{\perp}v \parallel_1 \\ \text{s. t.}\quad & \widehat{V}^{\mathrm{T}}v = 1 \end{aligned} \tag{4}$$

The grid search algorithm for the variance minimization problem (4) is given by the following algorithm.

**Algorithm 2. 1** The variance minimization approach via grid search algorithm

1. Sort the columns of $X^{(j)}$ in descending order of variance. Then the first variable has the maximum variance value, and its corresponding loading vector $v=(1,0,\cdots,0)$ is used as the first approximation of the solution, where the length of $v$ is $p-j+1$.

2. For $l=1,2,\cdots,m$, using the iterative steps to update all the components of the vector $v$: For $1\leqslant i\leqslant p-j+1$, update the $i$-th component $v^i$ of the current best approximation $v$, which is achieved by maximizing the following equation to find $\gamma^*$:

$$f(v^1 b(\gamma),\cdots,v^{i-1}b(\gamma),\cos\gamma,v^{i+1}b(\gamma),\cdots,v^{p-j+1}b(\gamma)),$$

where $\gamma$ takes the value in the interval $[\arccos(v^i-\pi/2^{l-1}), \arccos(v^i+\pi/2^{1-1})]$, $b(\gamma)=\sin(\gamma)/\sqrt{1-(v^i)^2}$ makes such unit standard conditions hold. This function is maximized by using grid search. The updated value of $v^i$ is $\cos\gamma^*$.

Note that if the number of iterations $l$ increases, we will implement a more rigorous search on the plane, because we assume that it is very close to the solution. Since $N$ grid is continuous, we can increase the accuracy in each iteration step. When the absolute value of the change of the optimal direction in the two iterations is lower than the predetermined tolerance level, we call the algorithm converges. When the maximum number of iterations is reached, the algorithm stops. Finally, the optimal sparse direction of the $j$-th principal component obtained by the grid algorithm must be rotated to the original space, i. e. $v_j=V_{j-1}^{\perp}v$.

## 2. 2 Reconstruction error minimization (REM)

While the VM approach obtains the PCs by maximizing the variance, we can also derive the PCs via minimizing the distance between projected data and the original ones. As described in Reference [16], this method can be formulated as follows,

$$\begin{aligned} \min_{v} & \sum_{i=1}^{n} \parallel x_i - VV^{\mathrm{T}}x_i \parallel^2 \\ \text{s. t.}\quad & V^{\mathrm{T}}V = I_k \end{aligned} \tag{5}$$

Inspired by this formulation, Zou et al.[12] reconstructed the product of the loading matrix, $V^{\mathrm{T}}V$, into two $p \times k$ matrices $A^{\mathrm{T}}B$. Based on this reconstruction, they proposed a sparse PCA method by adding a combination of $l_1$ and $l_2$ penalties to $B$ and leaving the orthogonal constraint to $A$, that is,

$$\begin{aligned} \min_{A,B} & \sum_{i=1}^{n} \parallel x_i - AB^{\mathrm{T}}x_i \parallel^2 + \\ & \lambda\sum_{j=1}^{k} \parallel B_j \parallel^2 + \sum_{j=1}^{k}\lambda_{1,j} \parallel \beta_j \parallel_1 \\ \text{s. t.}\quad & A^{\mathrm{T}}A = I_k \end{aligned} \tag{6}$$

where $\beta_j$ is the $j$-th column of $B$, $\lambda$ is a parameter controlling the norm of loading vectors, and $\lambda_{1,j}$ is a parameter that controls the sparsity of the load vector $\beta_j$. For data set with $n>p$, assume that $\lambda>0$, and for data set with $n\leqslant p$, $\lambda=0$.

The REM approach links the estimation of sparse PCs with variable selection in linear regression. When fixing the matrix $A$, solving the problem with respect to $B$ is an elastic net problem, which has be well studied, see Reference [17] for example. Based on this, Zou et al.[12] developed the block coordinate descent algorithm to the REM approach by optimizing the variables of (5), i. e., $A$ and $B$, in two separate sub-problems. The detail algorithm is presented as follows.

**Algorithm 2. 2** The REM approach via bolck coordinate descent method

1. Let $A$ be initialized to $V[,1:k]$, which is the load of the first $k$ ordinary principal components.

2. For a given $A=[\alpha_1,\cdots,\alpha_k]$, for $j=1,2,\cdots,k$, solve the following elastic network problem:

$\beta_j = \arg\min\limits_{\beta}(\alpha_j - \beta)^{\mathrm{T}} X^{\mathrm{T}} X(\alpha_j - \beta) + \lambda \|\beta\|^2 + \lambda_{1,j} \|\beta\|_1$.

3. For a given $B = [\beta_1, \cdots, \beta_k]$, calculate the SVD decomposition $X^{\mathrm{T}}XB = UDV^{\mathrm{T}}$, and then update $A = UV^{\mathrm{T}}$.

4. Repeat Steps 2,3 until convergence.

5. Standardization: $\widehat{V}_j = \beta_j / (\|\beta_j\|), j = 1, \cdots, k$.

As described in Reference [12], the empirical evidence shows that the output result of the algorithm does not change much with the change of $\lambda$. For the case of $n>p$, the default term $\lambda$ can be zero. In fact, $\lambda$ usually takes a small positive number to overcome the potential collinearity problem of $\lambda$.

### 2.3 Singular value decomposition (SVD)

An alternative way to obtain the loading matrix product $V^{\mathrm{T}}V$ is by using the singular value decomposition (SVD). Mathematically, let the SVD of $X$ be $X = UDV^{\mathrm{T}}$, where $U = [u_1, \cdots, u_r]$ with orthogonal columns, $V = [v_1, \cdots, v_r]$ with orthogonal columns, and $D = \mathrm{diag}\{d_1, \cdots, d_r\}$ with $d_1 \geqslant \cdots \geqslant d_r$. It can be easily to see that the $j$-th loading vector of $X^{\mathrm{T}}X$ is $v_j$, $j = 1, \cdots, r$. Thus we can obtain the PCs of $X^{\mathrm{T}}X$ by performing SVD to matrix $X$.

Since the SVD of $X$ can be rewritten as $\sum\limits_{j=1}^{r} d_j u_j v_j^{\mathrm{T}}$, we can estimate the singular vectors $u_j$ and $v_j$ one by one. In specific, let $\tilde{u}$ be an $n$-dimensional vector with unit norm, and $\tilde{v}$ be a $p$-dimensional vector. Then by the low-rank approximation property of SVD, we have the solution of the following optimization problem:

$$\min_{\tilde{u}, \tilde{v}} \|X - \tilde{u}\tilde{v}^{\mathrm{T}}\|_F^2 \tag{7}$$

is

$$\tilde{u} = u_1, \tilde{v} = d_1 v_1.$$

Similarly, we can derive the $j$-th components $(u_j, d_j v_j)$ by finding the best rank-1 approximation of the residual matrix $X - \sum\limits_{l=1}^{j-1} d_l u_l v_l^{\mathrm{T}}$.

Shen and Huang[11] introduced sparsity promoting penalty into (7) to obtain a sparse version of $\tilde{v}$,

$$\min_{\tilde{u}, \tilde{v}} \{ \|X - \tilde{u}\tilde{v}\|_F^2 + P_\lambda(\tilde{v}) \} \tag{8}$$

where $P_\lambda(\tilde{v}) = \sum\limits_{j=1}^{p} p_\lambda(|\tilde{v}_j|)$ is a penalty function with the parameter $\lambda$. To solve the problem (8), an iterative algorithm is developed. Firstly, given $\tilde{v}$, the solution of problem (8) with respect to $\tilde{v}$ has the explicit expression, i. e., $\tilde{u} = X\tilde{v} / \|X\tilde{v}\|$ [11]. Next, we discuss the optimization problem of $\tilde{v}$ with fixed $\tilde{u}$. Since $P_\lambda(\tilde{v}) = \sum\limits_{j=1}^{p} p_\lambda(|\tilde{v}_j|)$, the objective function in (8) can be rewritten as

$$\sum_i \sum_j (x_{ij} - \tilde{u}_i \tilde{v}_j)^2 + \sum_j p_\lambda(|\tilde{v}_j|) = \\ \sum_j \{ \sum_i (x_{ij} - \tilde{u}_i \tilde{v}_j)^2 + p_\lambda(|\tilde{v}_j|) \} \tag{9}$$

in which we can optimize each component of $\tilde{u}$ separately. For the $j$-th component $\tilde{v}_j$, we only need to solve the following problem:

$$\min_{\tilde{v}_j} \{ \tilde{v}_j^2 - 2(X^{\mathrm{T}}\tilde{u})_j \tilde{v}_j + p_\lambda(|\tilde{v}_j|) \}.$$

The above discussion are summarized as follows.

**Algorithm 2.3** The SVD approach via iterative method

1. Initialization: Apply standard SVD to $X$, and then get the optimal rank 1 of $X$ approximately as $su^* v^{*\mathrm{T}}$, where $u^*$ and $v^*$ are all unit vectors. Let $\tilde{v}_{\mathrm{old}} = sv^*$ and $\tilde{u}_{\mathrm{old}} = u^*$.

2. Update:

   (a) The element $\tilde{v}_j$ of $\tilde{u}_{\mathrm{new}}$ is $\min\limits_{\tilde{v}} \{ \tilde{v}_j^2 - 2(X^{\mathrm{T}}\tilde{u})_j \tilde{v}_j + p_\lambda(|\tilde{v}_j|) \}$;

   (b) $\tilde{u}_{\mathrm{new}} = X\tilde{v}_{\mathrm{new}} / \|X\tilde{v}_{\mathrm{new}}\|$.

3. Repeat Step 2 until convergence.

4. Standardize the last $\tilde{v}_{\mathrm{new}}$ to get the required sparse load vector $v = \tilde{v}_{\mathrm{new}} / \|\tilde{v}_{\mathrm{new}}\|$.

It can be seen that Algorithm 2.3 only contains simple linear iteration and group reduction rules. Therefore, it has the advantages of easy implementation and high computational efficiency.

### 2.4 Probabilistic model (PM)

Sigg and Buhmann[14] showed that the PCs can also be redefined as the maximum likelihood solution of a probabilistic latent variable model. The original data is assumed to be

$$X = ZW + \varepsilon \tag{10}$$

where $Z = \{Z_1, \cdots, Z_n\}^{\mathrm{T}}$ is the latent variable, $W \in \mathbb{R}^{k \times p}$ represents the principal component, its row vector is regarded as a $k$-dimensional latent variable, $\epsilon$ is the noise. Both the latent variable and noise are assumed to follow the normal distribution, that is, $Z_i \sim N(0, I_k)$, $\varepsilon \sim N(0, \sigma^2 I_p)$. Then, the marginal distribution of $X$ is a normal distribution with a mean value of 0 and a variance of $W^{\mathrm{T}}W + \sigma^2 I_p$. The estimation of the parameters $W$ and $\sigma^2$ can be achieved by the maximum likelihood function. Sigg and Buhmann[14] proposed using the variational maximum expectation algorithm to estimate the parameters.

To obtain sparse PCs, we add a step of axis-aligned gradient descent to the probabilistic model, and the detailed algorithm is presented as follows:

**Algorithm 2.4** The PM approach via gradient descent method

1. Initialize $t = 1$, apply standard SVD procedure to $X$, and then get the first principal component of $X$ as $W_{(t)}$.

2. When $|W_{(t+1)}^{\mathrm{T}} W_{(t)}| > 1 - \varepsilon$, the following loop is performed:

   (1) $y = XW_{(t)}$;

   (2) $W^* = \dfrac{\sum_{n=1}^{N} y_{(n)} x_{(n)}}{\sum_{n=1}^{N} y_{(n)}^2}$;

   (3) assign a value to $s$, its elements are $|w_i^*|$ in descending order, and the order is recorded in $\pi$;

   (4) for $k = 1, \cdots, K$, add $(s_k - s_{k+1})$ to the $1, \cdots, k$ elements of

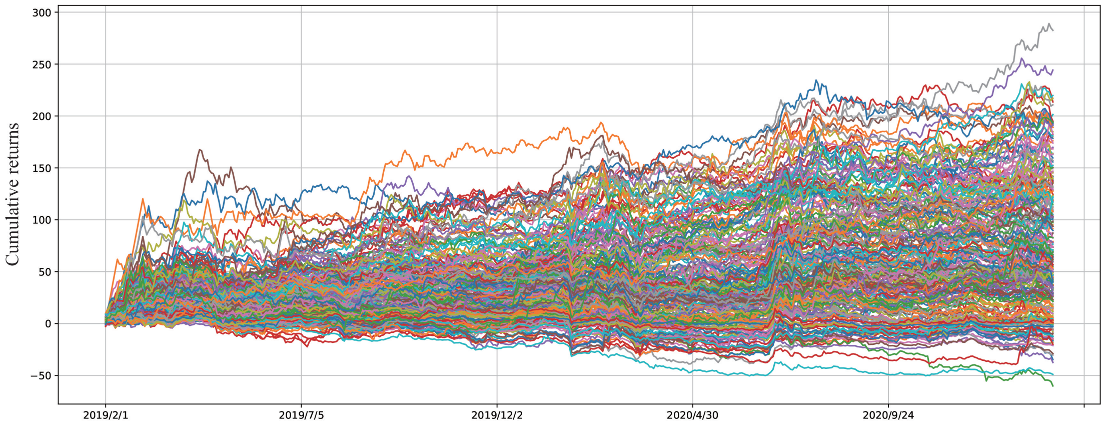


**Figure 2.** Cumulative returns of the stocks in the CSI 300 index from February 1, 2019 to February 1, 2021.

$X_{(t+1)}$;

(5) rearrange the elements in $W_{(t+1)}$ according to $\pi$, $t=t+1$.

3. Output $W$.

# 3 Empirical analysis

## 3.1 Data exhibition

We investigate the performance of the classic PCA and the sparse PCA methods described in Section 2 to the stock data from February 1, 2019 to February 1, 2021. We obtained the daily yield of the 234 unique stocks in the CSI 300 index from the Wind database. Figure 2 shows the cumulative daily return of 234 stocks in CSI 300 since February 1, 2019. It can be seen that some stocks are showing a clear upward trend, while others are showing a downward trend.

We define the sectors as General Industry Classification groups (GICs)[18], see Table 2.

## 3.2 Results from PCA

We first divide the whole period into two non-overlap periods: ① the period before the outbreak of the COVID-19 pandemic: from February 1, 2019 to December 31, 2019; ②the period after the outbreak: from January 1, 2020 to February 1, 2021.

**Table 2.** General Industry Classification (GIC) sectors.

| GIC($k$) | Description | Number of companies |
|---|---|---|
| 1 | Consumer discretionary | 16 |
| 2 | Consumer staples | 18 |
| 3 | Energy | 6 |
| 4 | Financials | 49 |
| 5 | Industrials | 37 |
| 6 | Information technology | 25 |
| 7 | Health care | 25 |
| 8 | Materials | 11 |
| 9 | Real estate | 16 |
| 10 | Telecommunication services | 7 |
| 11 | Utilities | 23 |

Figure 3 shows the first five principal components calculated from PCA before and after the outbreak. It can be seen that in 2019, the first PC has a large variance, about 37.4%, followed by 5.2%, 4.8%, 2.3% and 2.2%.

In comparison, after the outbreak, there is an obvious increment in the variance value of the second PC, while the variance of the first PC drops down to 33.6%. It suggests that the second PC absorbed some variations from the first PC after the outbreak. To provide further insight into the impact of the pandemic, let us look at the factor loading graph on the right panels in Figure 3. While the first PC has a similar coefficient values in most stocks, yet the directions of factor loadings in the second PC change substantially after the outbreak.

As known, the first PC represents a linear combination of input data that explains most of the differences. It turns out that this is the “market factor”, i.e., the trend of securities to rise and fall together as an asset class. The right panel of Figure 3 illustrates this by showing that all stocks have the same sign on the first PC. In other words, it is empirically the case that there is a dominant systematic factor called the equity risk premium explaining the variance of returns. This is because macro variables, such as monetary, fiscal policy, growth expectations, political risk, regulatory risk and other factors, influence the returns of all stocks.

Figure 4 plots the return of the investment portfolio based on the first and second principal components and the CSI 300 index. It can be seen that both the portfolios from the first and second PCs have

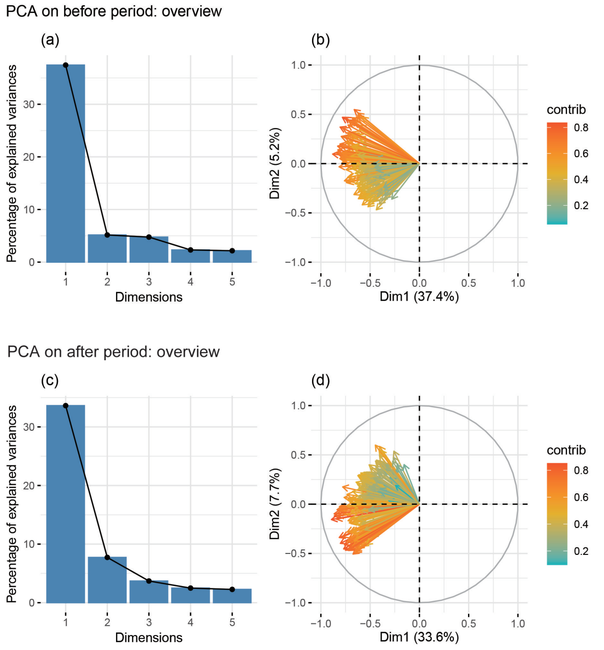


**Figure 3.** The first five PCs from classic PCA before and after the outbreak of the COVID-19 pandemic.

higher cumulative returns than the CSI 300 index. Furthermore, the portfolio returns of these two PCs are similar in 2019, but after January 1, 2020, the total returns of the portfolio of the second PC is significantly higher than those of the first PC. This might be due to that the sudden outbreak led to high demand in certain areas, which resulted in the growth of stocks in those fields. We will explore this furthermore via sparse PCA technique.

Figure 5 shows the daily profit by zooming in the top left panel of Figure 4 during the period of January 1, 2020 and April 20, 2020. In addition, the number of new domestic cases is included for comparison. We divide this period into five stages according to the portfolio changes.

(Ⅰ) The first stage (before January 9, 2020): The stock price fluctuates steadily, and there is no obvious upward or downward trend. At this time, news of "unexplained pneumonia" was only spread in a small area and had not attract the attention of the government and the public, so it had no obvious impact on the stock market.

(Ⅱ) The second stage (January 9, 2020 to February 3, 2020): On January 9, the expert evaluation team of the National Health Commission released information on the pathogen of Wuhan's unexplained virus pneumonia and determined that the pathogen was a new type of coronavirus. Based on the fear of SARS, a coronavirus in 2003, the public panicked and began to stock up large amounts of medical supplies such as masks and disinfectants. On January 23, 2020, Wuhan and several cities in Hubei issued the "lockdown" rule. At the time of the Lunar New Year, the population movement was large, which caused the pandemic to spread to a certain extent across the country. The entire Spring Festival is the most severe period of the pandemic, and the surge in the number of confirmed cases continues to challenge the public confidence. Since the market is closed during the Spring Festival, the CSI 300 index fell by 8% on the first day of opening after the Spring Festival. Although the pandemic did not spread across the

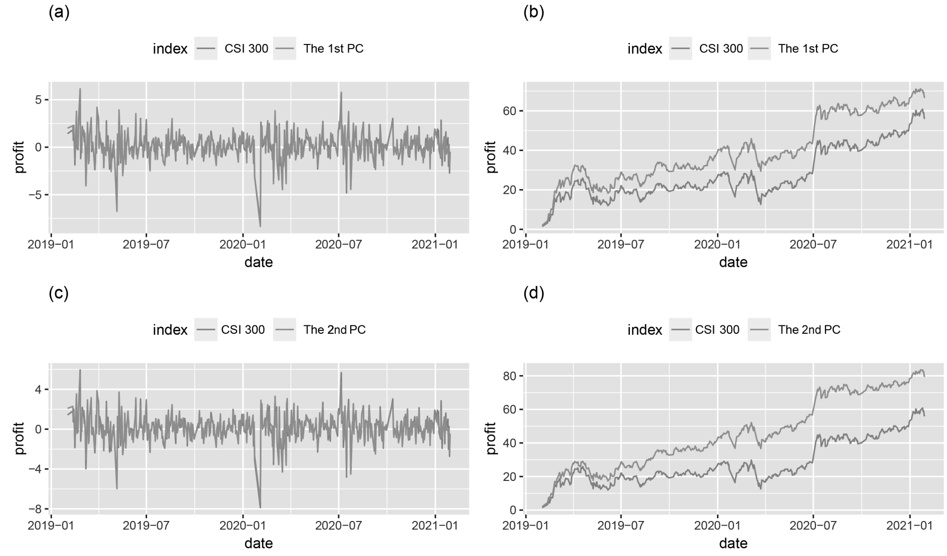


**Figure 4.** Portfolio for the stocks from the first two PCs. The CSI 300 index is also included for comparison. The daily and cumulative portfolios are presented in the left and right panels, respectively.

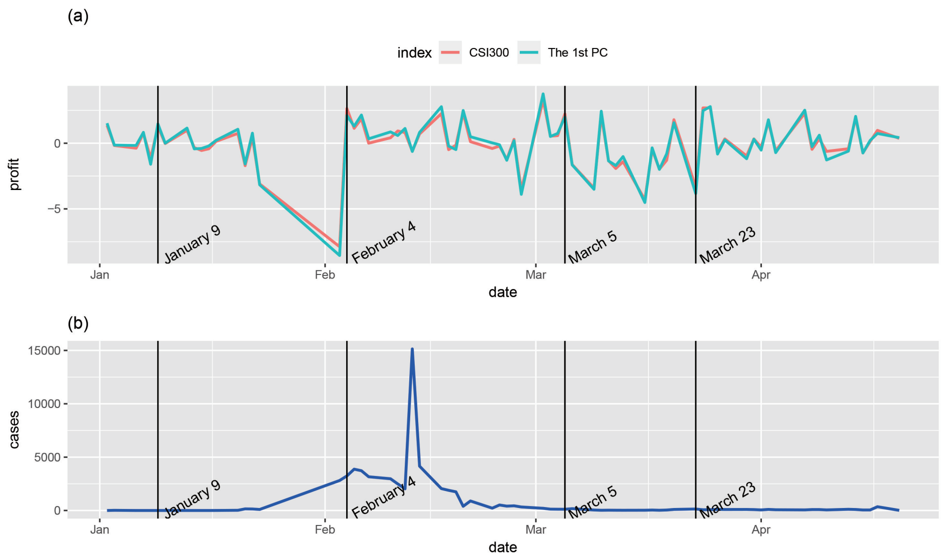


**Figure 5.** Portfolio for the stocks from the first PC and the CSI 300 index from January 1, 2020 to April 20, 2020. The number of new domestic cases is included in the bottom panel. The vertical black line indicates the division of five stages, which are defined by the portfolio changes.

country, the CSI 300 index showed a volatile decline.

(Ⅲ) The third stage (February 4, 2020 to March 5, 2020): On February 3, Wuhan Huoshenshan Hospital received the first batch of patients. The P3 Laboratory in Zhejiang Province has isolated 8 strains of viruses, several of which are very suitable for vaccines. At the same time, with the unfolding of the anti-pandemic, the number of newly diagnosed patients has begun to decline, and public confidence has been greatly improved. The stock market has shown an upward trend amidst volatility.

(Ⅳ) The fourth stage (March 5, 2020 to March 23, 2020): On March, the pandemic in China was basically under control, but overseas pandemic s began to break out, and U. S. stocks experienced four circuit breakers. Affected by the U. S. stock market, the domestic stock market experienced a major decline.

(Ⅴ) The fifth stage (March 23, 2020 to April 20, 2020): In the second half of March, domestic confirmed diagnoses are basically cleared. With the resumption of work and production on a large scale, the stock market began to pick up and gradually returned to the average line before the pandemic.

(Ⅵ) The sixth stage (after April 20, 2020): Starting from the second half of April, the domestic pandemic has been well controlled, and the order of production and life before the pandemic has been completely restored. The stock market has also returned to a state of steady volatility before the pandemic. However, after experiencing such a rapid and powerful pandemic, the structure of the stock market in the post-pandemic era requires further analysis.

### 3.3 Results from the sparse PCA methods

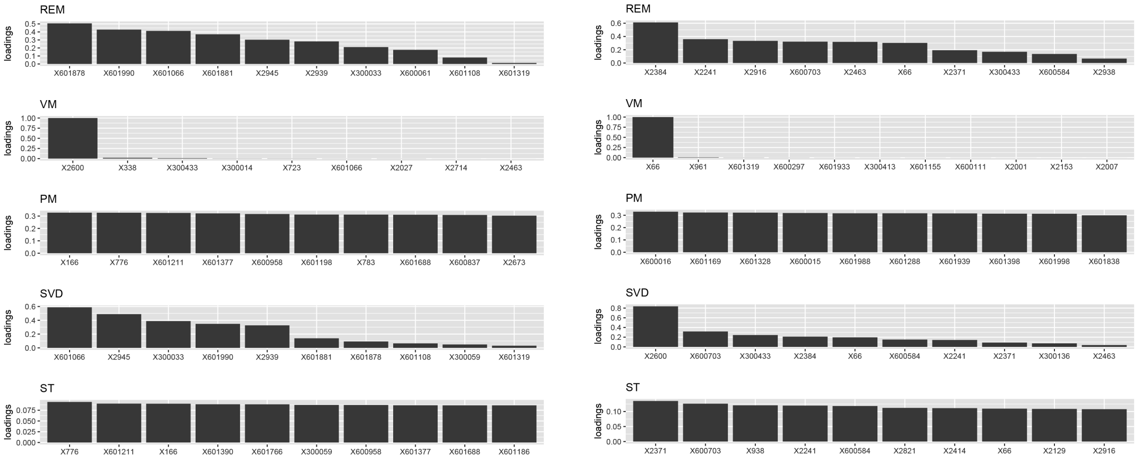
The results from PCA shows how the pandemic influences the stock market. Next we will present more difference before and after the outbreak, especially the difference in the leading stocks set. We fix the number of the first and second PCs to be 10, and perform four sparse PCA methods mentioned in Section 2 on the stock data. We also include the simple truncating (ST)[2] method by keeping the top 10 elements of PCs (in absolute value) from the classic PCA method and truncating the remaining elements to be zero.

First of all, we applied all the above methods to the entire period. Table 3 shows the GIC of the nonzero elements in the first two sparse PCs. It can be seen that the stocks selected by the first PC mainly belong to the financial field, while those of the second PC mainly belong to the health care or industrial field except for the PM approach. We will provide further insight into it by separating the period by the outbreak.

Figure 6 and Figure 7 present the selected stocks in the first two PCs for the periods before and after the outbreak, respectively. The stocks are arranged according to the absolute value of the loadings. We also include the GIC sector that the stocks belong to in Table 4 and Table 5. It can be seen that before the outbreak, the stocks in the first PC are mostly from financial field for all methods except the VM method. Different from other methods, both the two PCs from VM have complicate constructions. For the second PC, both REM and SVD yields to a combination of industrial stocks. The second PC of PM methods still consists of financial stocks.

**Table 3.** The nonzero loading fields from different sparse PCA techniques on entire period.

| | Rank | REM | VM | PM | SVD | ST |
|---|---|---|---|---|---|---|
| 1st PC | 1 | Information technology | Materials | Financials | Financials | Financials |
| | 2 | Financials | Industrials consumer | Financials | Information technology | Financials |
| | 3 | Financials | Discretionary | Financials | Financials | Financials |
| | 4 | Information technology | Consumer staples | Financials | Financials | Financials |
| | 5 | Financials | Industrials | Financials | Information technology | Financials |
| | 6 | Financials | Consumer staples | Financials | Financials | Financials |
| | 7 | Financials | Industrials consumer | Financials | Financials | Financials |
| | 8 | Energy | Discretionary information | Financials | Financials | Financials |
| | 9 | Financials | Technology | Financials | Financials | Financials |
| | 10 | Financials | Industrials | Financials | Financials | Information technology |
| 2nd PC | 1 | Health care | Industrials | Financials | Industrials | Financials |
| | 2 | Health care | Industrials | Financials | Industrials | Financials |
| | 3 | Health care | Industrials technology | Financials | Information | Financials |
| | 4 | Health care | Health care | Financials | Industrials | Financials |
| | 5 | Materials | Industrials | Financials | Industrials | Financials |
| | 6 | Health care | Industrials | Financials | Industrials | Financials |
| | 7 | Health care | Consumer discretionary | Financials | Industrials | Financials |
| | 8 | Health care | Consumer staples | Financials | Industrials | Financials |
| | 9 | Health care | Industrials | Financials | Industrials consumer | Financials |
| | 10 | Health care | Consumer staples | Financials | Consumer discretionary | Financials |



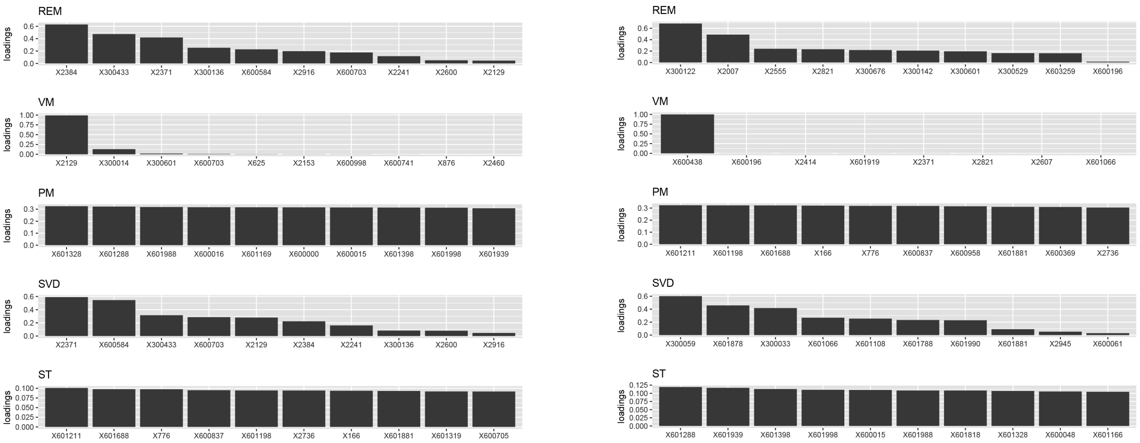
（a） The 1st PC loading

（b） The 2nd PC loading

**Figure 6.** The selected stocks in the first two PCs for the periods before the outbreak of the COVID-19 pandemic. The stocks are arranged according to the absolute value of the loadings.

After the outbreak, the selected stocks are totally different, especially for the REM and VM methods. For REM and VM, stocks from the health care field are identified to contribute to the first two PCs, while the stocks from financial and real estate fields disappear. This might be because of the urgent need of medical supplies during the period of the COVID-19 pandemic.

It can be seen that the first and second PC of ST algorithm is close to the PM algorithm.



（a） The 1st PC loading

（b） The 2nd PC loading

**Figure 7.** The selected stocks in the first two PCs for the periods after the outbreak. The stocks are arranged according to the absolute value of the loadings.

**Table 4.** The nonzero loading fields of of the first two sparse PCs ( before the outbreak ).

| | Rank | REM | VM | PM | SVD | ST |
|---|---|---|---|---|---|---|
| 1st PC | 1 | Financials | Industrials | Financials | Financials | Financials |
| | 2 | Financials | Utilities | Financials | Financials | Financials |
| | 3 | Financials | Industrials | Financials | Information technology | Financials |
| | 4 | Financials | Materials | Financials | Financials | Real estate |
| | 5 | Financials | Energy | Financials | Financials | Utilities |
| | 6 | Financials | Financials | Financials | Financials | Information technology |
| | 7 | Information technology | Information technology | Financials | Financials | Financials |
| | 8 | Financials | Consumer discretionary | Financials | Financials | Financials |
| | 9 | Financials | Industrials | Financials | Information technology | Financials |
| | 10 | Financials | | Financials | Financials | Real estate |
| 2nd PC | 1 | Information technology | Information technology | Real estate | Industrials | Industrials |
| | 2 | Industrials | Real estate | Real estate | Industrials | Industrials |
| | 3 | Industrials | Financials | Real estate | Industrials | Information technology |
| | 4 | Industrials | Utilities | Real estate | Information technology | Industrials |
| | 5 | Industrials | Consumer staples | Real estate | Information technology | Industrials |
| | 6 | Information technology | Information technology | Utilities | Industrials | Health care |
| | 7 | Industrials | Real estate | Real estate | Industrials | Industrials |
| | 8 | Industrials | Materials | Materials | Industrials | Information technology |
| | 9 | Industrials | Health care | Real estate | Industrials | Consumer discretionary |
| | 10 | Industrials | Information technology | Materials | Industrials | Industrials |
]

**Table 5.** The nonzero loading fields of the first two sparse PCs (after the outbreak).

| | Rank | REM | VM | PM | SVD | ST |
|---|---|---|---|---|---|---|
| 1st PC | 1 | Information technology | Consumer discretionary | Financials | Industrials | Financials |
| | 2 | Industrials | Materials | Financials | Industrials | Financials |
| | 3 | Industrials | Health care | Financials | Industrials | Financials |
| | 4 | Industrials | Industrials | Financials | Industrials | Financials |
| | 5 | Industrials | Utilities | Financials | Consumer discretionary | Financials |
| | 6 | Industrials | Information technology | Financials | Information technology | Financials |
| | 7 | Industrials | Health care | Financials | Industrials | Financials |
| | 8 | Industrials | Utilities | Financials | Industrials | Financials |
| | 9 | Industrials | Consumer discretionary | Financials | Industrials | Financials |
| | 10 | Consumer discretionary | Materials | Financials | Industrials | Financials |
| 2nd PC | 1 | Health care | Consumer discretionary | Financials | Information technology | Financials |
| | 2 | Health care | Health care | Financials | Financials | Financials |
| | 3 | Information technology | Industrials | Financials | Information technology | Financials |
| | 4 | Health care | Utilities | Financials | Financials | Financials |
| | 5 | Health care | Industrials | Financials | Financials | Financials |
| | 6 | Health care | Health care | Financials | Financials | Financials |
| | 7 | Health care | Consumer discretionary | Financials | Financials | Financials |
| | 8 | Health care | Financials | Financials | Financials | Financials |
| | 9 | Health care | | Financials | Financials | Real estate |
| | 10 | Health care | | Financials | Financials | Financials |

In addition, according to the weights of sparse PCs, we can formulate a winning portfolio, which selects companies with nonzero loadings. As shown in Figure 8, the resulting portfolio will perform significantly better than the market because it invests in companies that have actually benefited from the pandemic. It can be seen that except for the PM and ST methods, all the other methods provide sparse PCs with better performance compared to the CSI 300 index. Regardless of whether it is formulated from the first or second PC, the return of PM and ST portfolio is lower than CSI 300 index. Based on the previous analysis, it may because that the stocks selected by PM and ST are mostly concentrated in bank stocks, which makes it difficult to diversify risks. The first PC of all methods has a similar trend to the CSI 300 index, which suggests that the first PC represents market risk. Compared with the first PC, the second PC of REM, VM and SVD portfolio are rising more steadily, surpassing most of the cumulative return of the first PC in the later period. In short, most of the sparse PCA methods yield better returns than the CSI 300 index, and the second PC has more robust and stable performance compare with the first PC. Combining the PCA results in Section 3.2, we can tell the reason that the cumulative returns of the second PC is higher than those of the first PC in the period after the outbreak. It is because the dominant stocks of the second PC are all from the health care and industrial fields, which have achieved rapid growth during the pandemic period.

Finally, Figure 9 shows the return trend of the portfolio obtained from the sparse PC during the period of January 1, 2020 and April 20, 2020. We include the results of the REM and SVD methods as example. Compared with the classic PCA approach, the return trends show more volatility between the sparse PCA derived portfolio and the CSI 300 index. In specific, the portfolio of the first PC tends to have consistent performance with the CSI 300 index with a smaller variance, which is expected since it consists of stocks

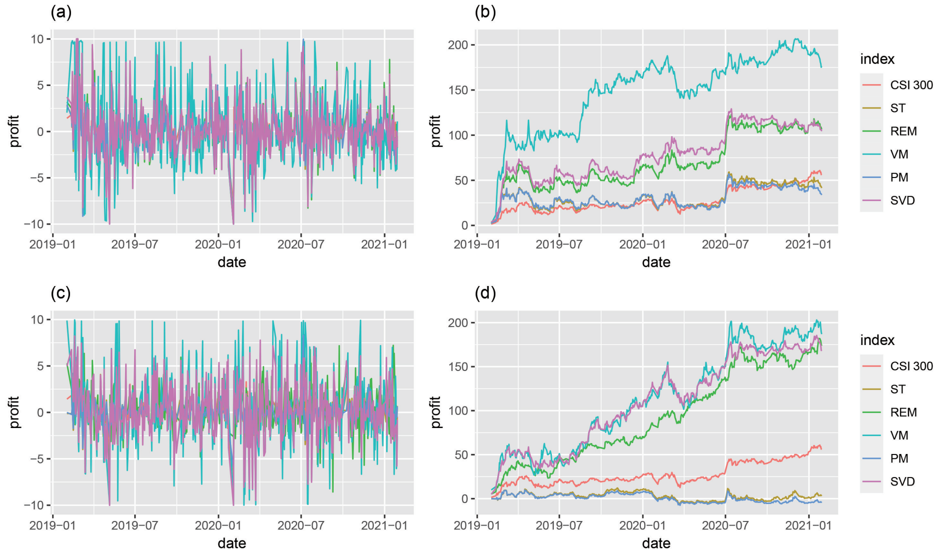


**Figure 8.** Portfolio for the stocks from the first two sparse PCs. The CSI 300 index is also included for comparison. The daily and cumulative portfolios are presented in the left and right panels， respectively.

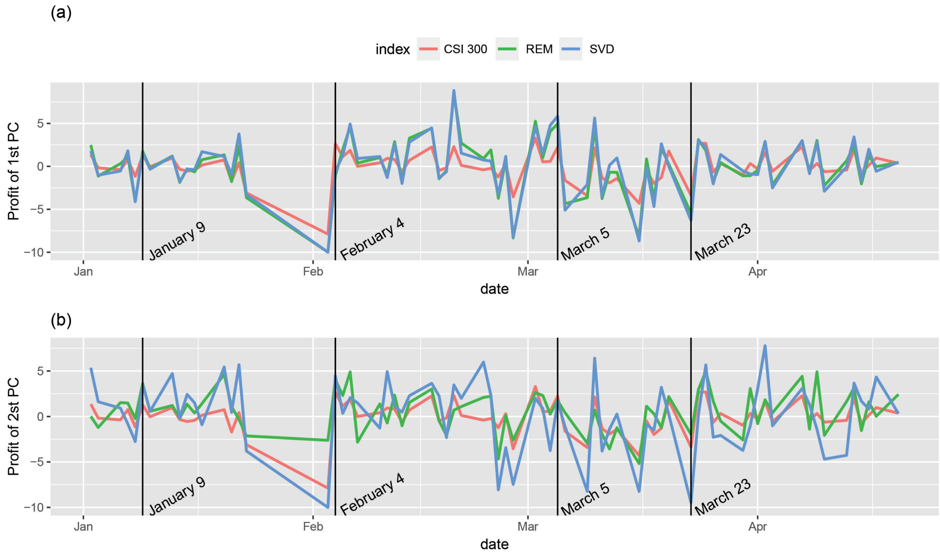


**Figure 9.** Portfolio for the stocks from the first two sparse PCs. The CSI 300 index is also included for comparison. The 1st PC and 2nd PC are presented in the upper and down panels， respectively.

from the financial sector. The portfolio of the second PC have a total different trend， which can be explained by its composition， i. e. ， stocks are all from the health care and industrial fields.

## 4　Numerical research

In this section， we compare the performance of various algorithms on synthetic data. All programs are completed by R software， and the corresponding sparse principal component analysis methods are： ①VM， using R package： pcaPP； ②REM， using R package： elasticnet； ③M， using R package ：Nsprcomp； ④SVD： Consider two algorithms： one is SVDb algorithm， the other is SVDc algorithm using R package： PMD.

### 4.1　Setting

In this section， we consider a more complex data generating mechanism by extending the studies in Hsu et al.[19]. Suppose we are given a sample covariance $\Sigma$ coming from a “spiked” model of covariance， with

$$\Sigma = VV^{\mathrm{T}} + \sigma^2 I_p$$

where the columns of $V=(v_{ij}) \in \mathbb{R}^{p\times 3}$ are the true sparse leading eigenvector， $I_p$ is an identity matrix， and $\sigma = 0.1$. A data matrix $X \in \mathbb{R}^{n\times p}$ is then generated by drawing $n = 100$ samples from a zero-mean normal distribution with covariance matrix $\Sigma$， that is， $X \sim$

$N(0,\Sigma)$.

Two dimension are considered: $p=50$ and $p=100$.

In addition, we consider two different covariance structures:

① Non-overlap structure. (Figure 10(a))

$$v_1 = (\underbrace{0.32,\cdots,0.32}_{10},\underbrace{0,\cdots,0}_{p-10}),$$

$$v_2 = (\underbrace{0,\cdots,0}_{10},\underbrace{0.32,\cdots,0.32}_{10},\underbrace{0,\cdots,0}_{p-20}),$$

$$v_3 = (\underbrace{0,\cdots,0}_{20},\underbrace{0.32,\cdots,0.32}_{10},\underbrace{0,\cdots,0}_{p-30}).$$

② Overlap structure. (Figure 10(b))

$$v_1 = (\underbrace{0.258,\cdots,0.258}_{15},\underbrace{0,\cdots,0}_{p-15}),$$

$$v_2 = (\underbrace{0,\cdots,0}_{5},0.1312,\underbrace{-0.061,\cdots,-0.061}_{3},0.061,\underbrace{0.022,\cdots,0.022}_{5},\underbrace{0.31,\cdots,0.31}_{10},\underbrace{0,\cdots,0}_{p-25}),$$

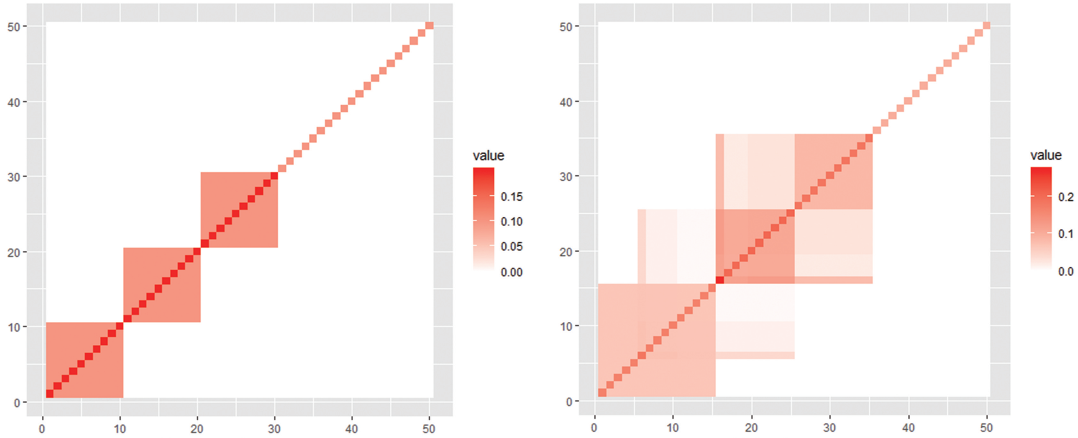
$$v_3 = (\underbrace{0,\cdots,0}_{15},\underbrace{-0.07,\cdots,-0.07}_{3},\underbrace{0.0994,\cdots,0.0994}_{6},\underbrace{0.292,\cdots,0.292}_{10},\underbrace{0,\cdots,0}_{p-35}).$$

Therefore, four simulation scenarios are produced, they are: ① non-overlapping covariance structure, the number of variables $p = 50$; ② non-overlapping covariance structure, the number of variables $p = 300$; ③ overlapping covariance structure, The number of variables $p = 50$; ④ Overlapping covariance structure, the number of variables $p=300$. The simulation in each case was repeated 300 times to estimate the first three principal components.

For comparison criteria, our point here is that, while variance versus cardinality is a direct way of comparing the performance of sparse PCA algorithms, accurate recovery of the support is often a far more important objective. Many methods produce similar variance levels given a limited budget of nonzero components, but their performance in recovering the true support is often markedly different. Therefore, we consider a new metric to evaluate the performance of the different methods, which is defined as the absolute value of the inner product between the true and estimated loading coeffcients, i. e, Cosine $= | v_j^{\mathrm{T}} \widehat{v_j} |$, $j = 1,2,3$.

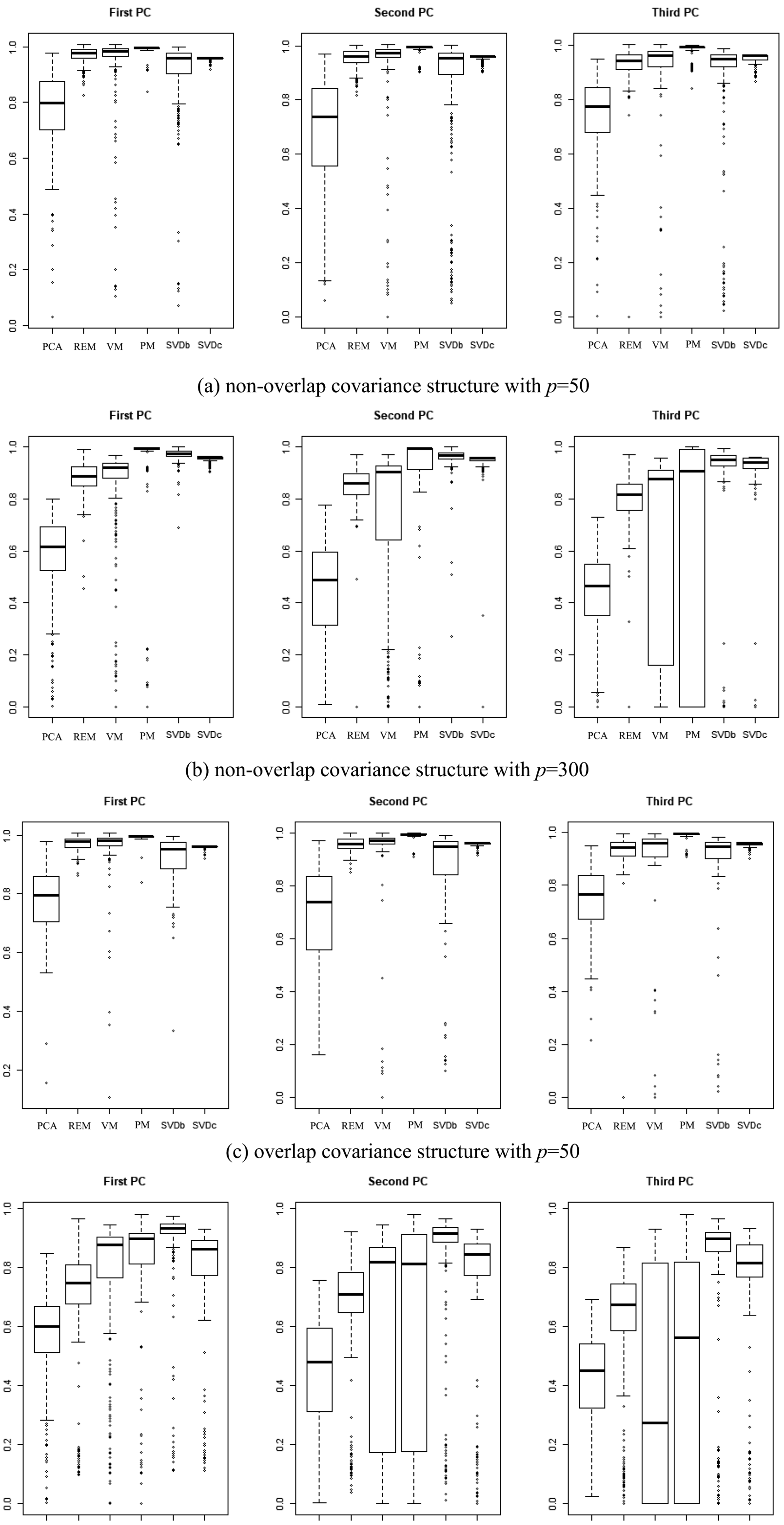
## 4.2 Results

The results of the four simulation studies are plotted in Figure 11. Obviously, almost all sparse PCA methods perform better than classic PCA methods. This shows that the sparse PCA method can better estimate the load coefficient value. Comparing the non-overlapping covariance structure and the overlapping covariance structure, the sparse PCA method performs better in the case of non-overlapping covariance structure (Figure 11(a)-(b) versus 11(c)-(d)). In addition, for the sparse PCA method, it can be observed that the high coincidence rate of PC1 exceeds that of PC2 and PC3. Also, sparse PCA performs better when the number of variables is small (Figure 11(a) versus 11(b) or 11(c) versus 11(d)). From the four pictures in Figure 11, we can find that the REM algorithm, SVDb algorithm (R code written by Shen and Huang[11]) and SVDc algorithm (R software package PMD) perform well



(a) Non-overlap structure (b) Overlap structure

**Figure 10.** An illustration of the covariance structure in synthetic examples.



(a) non-overlap covariance structure with $p$=50

(b) non-overlap covariance structure with $p$=300

(c) overlap covariance structure with $p$=50

(d) overlap covariance structure with $p$=300

**Figure 11.** Boxplots of the Consine values drawn by six algorithms：PCA, REM, VM, PM, SVDb and SVDc in four simulation schemes.

under any circumstances, while the VM algorithm and PM algorithm perform worse when the number of variables is greater than the number of samples ( $p>n$). This shows that the REM and SVD methods are more suitable for high-dimensional small sample data. On the other hand, it shows that the stability of these two methods is stronger than the latter two algorithms.

## 5 Conclusions

The COVID-19 pandemic has caused serious public health and economic consequences throughout the world. It is very important to assess the impact of the pandemic on the economy, especially the stock market. It is of great interest to study the impact of the pandemic on the stock market via dimension reduction techniques such as PCA. In practice, it is important to figure out which stocks are inflected mostly by the pandemic and construct portfolio management based on them. To this end, we collected the CSI 300 stock data from February 1, 2019 to February 1, 2021, and divided them into two periods, i. e., before and after January 1, 2020, and applied PCA and sparse PCA methods to them. The results show that the outbreak of the pandemic led to changes in both proportion and constitution of the principal component of the stocks in the CSI 300 index. In addition, our research work compares four different sparse PCA procedures in CSI 300 data as well as simulated data, which would provide a practical guidance for financial applicants.

Studies have shown that the performance of each algorithm of principal component analysis varies greatly, and there still have room for improvement. In the future study, it is of interest to use other modern optimization algorithms (such as primal-dual active set algorithm[20] and mixed integer optimization algorithm[21]) to derive a more accurate estimation. In addition, our current work is based on historical data before and after the outbreak of the COVID-19 pandemic, which can be used as a dynamic tool for index selection in the future, so we need to think about ways to improve its generalization ability.

## Acknowledgments

This work is partially supported by the NSFC (11801540), the Natural Science Foundation of Anhui (BJ2040170017) and the Science and Technology Program of Guangzhou (202002030129).

## Conflict of interest

The authors declare no conflict of interest.

## Author information

**LI Ming** is currently a master student under the supervision of Assoc. Prof. Wen Canhong in Department of Statistics and Finance, University of Science and Technology of China. Her research interests include high-dimensional statistics and machine learning.

**WEN Canhong** (corresponding author) received her PhD degree from Sun Yat-sen University. She is currently an Associate Professor at University of Science and Technology of China. Her research interests include statistical algorithm and learning, biostatistics and image genetics.

## References

[1] Yang L. An application of principal component analysis to stock portfolio management. Christchurch, New Zealand: University of Canterbury, 2015.

[2] Cadima J, Jolliffe I T. Loading and correlations in the interpretation of principle components. Journal of Applied Statistics, 1995, 22(2): 203-214.

[3] Vines S. Simple principal components. Journal of the Royal Statistical Society: Series C (Applied Statistics), 2000, 49 (4): 441-451.

[4] Ma Z. Sparse principal component analysis and iterative thresholding. Annals of Statistics, 2013, 41(2): 772-801.

[5] Jolliffe I T, Trendafilov N T, Uddin M. A modified principal component technique based on the LASSO. Journal of Computational and Graphical Statistics, 2003, 12 (3): 531-547.

[6] d'Aspremont A, El Ghaoui L, Jordan M I, et al. A direct formulation for sparse PCA using semidefinite programming. SIAM Review, 2007, 49(3): 434-448.

[7] Journée M, Nesterov Y, Richtárik P, et al. Generalized power method for sparse principal component analysis. Journal of Machine Learning Research, 2010, 11(2): 517-553.

[8] Moghaddam B, Weiss Y, Avidan S. Spectral bounds for sparse PCA: Exact and greedy algorithms. In: Proceedings of the 18th International Conference on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2005: 915-922.

[9] d'Aspremont A, Bach F, El Ghaoui L. Optimal solutions for sparse principal component analysis. Journal of Machine Learning Research, 2008, 9(7): 1269-1294.

[10] Croux C, Filzmoser P, Fritz H. Robust sparse principal component analysis. Technometrics, 2013, 55(2): 202-214.

[11] Shen H, Huang J Z. Sparse principal component analysis via regularized low rank matrix approximation. Journal of Multivariate Analysis, 2008, 99(6): 1015-1034.

[12] Zou H, Hastie T, Tibshirani R. Sparse principal component analysis. Journal of Computational and Graphical Statistics, 2006, 15(2): 265-286.

[13] Tipping M E, Bishop C M. Probabilistic principal component analysis. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 1999, 61 (3): 611-622.

[14] Sigg C D, Buhmann J M. Expectation-maximization for sparse and non-negative PCA. In: Proceedings of the 25th International Conference on Machine Learning. New York: Association for Computing Machinery, 2008: 960-967.

[15] Witten D M, Tibshirani R, Hastie T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics, 2009, 10(3): 515-534.

[16] Friedman J, Hastie T, Tibshirani R, et al. The Elements of Statistical Learning. New York: Springer, 2001.

(Continued on p.430)

we can obtain $\Pi=\frac{s(l-cl^2-s-(h(-1+ch)+s)\rho)}{2\rho}$. When $\frac{\partial \Pi}{\partial s}=\frac{l-cl^2+h(1-ch)\rho-2s(1+\rho)}{2\rho}=0$, the optimal price is $s^*=\frac{l-cl^2+h\rho-ch^2\rho}{2(1+\rho)}$. The second-order condition $\frac{\partial^2 \Pi}{\partial s^2}=-\frac{1+\rho}{\rho}<0$ is always satisfied.

### A. 6 Proof of Proposition 5. 1

The two first-order conditions are $\frac{\partial s^*}{\partial l}=\frac{1-2cl}{2+2\rho}$ and $\frac{\partial s^*}{\partial h}=\frac{\rho-2ch\rho}{2+2\rho}$. The two second-order conditions: $\frac{\partial^2 s^*}{\partial l^2}=-\frac{c}{1+\rho}<0$ and $\frac{\partial^2 s^*}{\partial h^2}=-\frac{c\rho}{1+\rho}<0$ are always satisfied.

### A. 7 Proof of Proposition 5. 2

Considering the calculation result is too complicated, this study only shows the calculation process. It can be known from $\Pi_L^*=\frac{(l(cl-1)+2h(ch-1)\rho)^2}{8\rho(4\rho+1)-4}$ and $\Pi^*=\frac{(l(cl-1)+h(ch-1)\rho)^2}{8\rho(1+\rho)}$:

(i) The platform's profit under the bidding mode is higher than that under the piece mode when $\frac{(l(cl-1)+h(ch-1)\rho)^2}{8\rho(1+\rho)}<\frac{(l(cl-1)+2h(ch-1)\rho)^2}{8\rho(4\rho+1)-4}$.

(ii) The platform's profit under the piece mode is higher than that under the bidding mode when $\frac{(l(cl-1)+h(ch-1)\rho)^2}{8\rho(1+\rho)}>\frac{(l(cl-1)+2h(ch-1)\rho)^2}{8\rho(4\rho+1)-4}$.

(Continued from p. 418)

[17] Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2005, 67(2): 301-320.

[18] Avellaneda M. Hierarchical PCA and applications to portfolio management. https://ssrn.com/abstract=3467712.

[19] Hsu Y L, Huang P Y, Chen D T. Sparse principal component analysis in cancer research. Translational Cancer Research, 2014, 3(3): 182-190.

[20] Wen C, Zhang A, Quan S, et al. BeSS: An R package for best subset selection in linear, logistic and CoxPH models. Journal of Statistical Software, 2020, 94(1): 1-24.

[21] Bertsimas D, Cory-Wright R, Pauphilet J. Solving large-scale sparse PCA to certifiable (near) optimality. https://arxiv.org/abs/2005.05195.

# 通过稀疏 PCA 分析新冠疫情对股市的影响

黎明,温灿红*

中国科学技术大学管理学院统计与金融系,安徽合肥 230026

*通讯作者. E-mail:wench@ustc.edu.cn

**摘要:** 新冠疫情的爆发在全世界造成了严重的公共卫生和经济后果。评估新冠疫情对经济,尤其是股市的影响非常重要。为此,我们提出应用几种最先进的稀疏主成分分析(PCA)方法来分析 2019 年 2 月 1 日至 2021 年 2 月 1 日的沪深 300 指数股票数据,以揭示新冠疫情爆发的影响. 将这段时间分为两个时期——2020 年 1 月 1 日之前和之后,在此基础上,我们尝试提取主成分并构建投资组合. 结果表明,在新冠疫情爆发之后,代表市场的主成分的比例有所下降. 关于前两个主成分的构成,新冠疫情爆发后,起决定作用的股票集合有很大的不同. 在新冠疫情之后,医疗保健行业的股票开始在沪深 300 指数的投资组合中发挥重要作用. 与沪深 300 指数相比,稀疏 PCA 方法的前两个主成分可以在组成投资组合的股票集数量少得多的情况下获得更高的回报. 综上所述,新冠疫情的爆发导致沪深 300 指数股票的主成分比例和构成发生了变化.

**关键词:** 新冠疫情;稀疏主成分分析;股票指数