PCA Analysis: Detailed Explanation

Coursework Imperial 2024

1. Data Matrix Z

Let $\mathbf{Z}$ be a matrix of standardized returns, where:

- T is the number of observations.
- N is the number of assets.

The matrix $\mathbf{Z}$ is of size $T \times N$:

$$\mathbf{Z} = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1N} \\ z_{21} & z_{22} & \cdots & z_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ z_{T1} & z_{T2} & \cdots & z_{TN} \end{bmatrix}$$

2. Compute the Correlation Matrix C

Calculate the correlation matrix $\mathbf{C}$, which is an $N \times N$ matrix:

$$\mathbf{C} = \frac{1}{T-1}\mathbf{Z}^T\mathbf{Z}$$

3. Eigenvalue Decomposition

Perform eigenvalue decomposition on the correlation matrix $\mathbf{C}$:

$$\mathbf{C} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$$

where:

- $\mathbf{V}$ is an $N \times N$ matrix whose columns are the eigenvectors of $\mathbf{C}$.
- $\mathbf{\Lambda}$ is an $N \times N$ diagonal matrix whose diagonal elements $\lambda_1, \lambda_2, \ldots, \lambda_m$ are the eigenvalues of $\mathbf{C}$.

4. Select the 20 Largest Eigenvalues

Sort the eigenvalues in descending order and select the first 20 eigenvalues. Let these eigenvalues be $\lambda_1, \lambda_2, \ldots, \lambda_{20}$.

5. Extract Eigenvectors Corresponding to the 20 Largest Eigenvalues

Select the eigenvectors corresponding to the 20 largest eigenvalues. These eigenvectors form the matrix $\mathbf{V}_{20}$, where:

$$\mathbf{V}_{20} = \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_{20} \end{bmatrix}$$

Here, $\mathbf{V}_{20}$ is an $N \times 20$ matrix.

6. Constructing the Hidden Factors

Using the selected eigenvectors and eigenvalues, we can construct the hidden factors.

For each of the top eigenvalues, we construct the hidden factors as follows:

$$F_j^t = \sum_{i=1}^{N} \frac{\lambda_{i,j}}{\sigma_i} R_i^t$$

where:

- $\lambda_{i,j}$ is the i-th component of the j-th eigenvector (ranked by eigenvalue magnitude).
- σ_i is the standard deviation of the i-th asset.
- R_i^t is the return of the i-th asset at time t.

For instance, if $\mathbf{v}_1$ is the first eigenvector, the one that corresponds to the largest eigenvalue. $\mathbf{v}_1$ is a vector of dimension N that we can write as:

$$\mathbf{v}_1 = (\lambda_{1,1}, \lambda_{2,1}, \cdots, \lambda_{N,1})$$

Then the first hidden factor at time t, F_1^t, is:

$$F_1^t = \sum_{i=1}^{N} \frac{\lambda_{i,1}}{\sigma_i} R_i^t$$

Thus, the hidden factor is a vector of dimension T (number of time observations). The matrix of hidden factors, considering the top 20 eigenvalues, has the dimension $(T, 20)$.

In summary, the process involves scaling the components of each eigenvector by the inverse of the standard deviation of the corresponding asset returns and then summing these scaled components, weighted by the returns, to form the hidden factors.