In a market consisting of N stocks, we denote the dividend-adjusted return on stock i at trading day t by $r_{i,t}$. We adopt a factor model for stock return,

$$\mathbf{r}_t - r_t^f \mathbf{1}_N = \mathbf{B}_t \mathbf{f}_t + \boldsymbol{\epsilon}_t, \quad t = 1, 2, \ldots, T \tag{1}$$

Here, $\mathbf{r}_t = \{r_{i,t}\}_{i=1}^N \in \mathbb{R}^N$ are the dividend-adjusted daily return, $r_t^f \in \mathbb{R}$ is the risk-free rate, $\mathbf{f}_t \in \mathbb{R}^{K\times 1}$ are the underlying factors, $\mathbf{B}_t \in \mathbb{R}^{N\times K}$ are the corresponding loadings on K factors, and $\boldsymbol{\epsilon}_t \in \mathbb{R}^N$ are the residual returns. Factor candidates varies widely, ranging from economical-driven factors such as the Fama-French factors, to statistically-driven factors derived from PCA. In our approach, factors are selected as the leading eigenvectors in PCA. The number of factors K is chosen based on the eigenvalue spectrum of the empirical correlation of daily returns.
Without loss of generality, these factors can be interpreted as portfolios of stocks,

$$F_t = \omega_t \left(r_t - r_f\right) \tag{2}$$

where $\omega_t \in \mathbb{R}^{K\times N}$ contains corresponding portfolio weights. Combining eq. (1) and eq. (2) yields

$$r_t - r_f = \beta_t \omega_t \left(r_t - r_f\right) + \epsilon_t \Rightarrow \epsilon_t = \left(I - \beta_t \omega_t\right)\left(r_t - r_f\right) := \Phi_t \left(r_t - r_f\right) \tag{3}$$

Here,

$$\Phi_t := \left(I - \beta_t \omega_t\right) \tag{4}$$

defines a linear transformation from r_t to ϵ_t. More importantly, $\epsilon_{i,t}$ can be viewed as the return of a tradable portfolio with weights specified by the i-th row of Φ_t. Consequently, the investing universe spanned by r_t is termed as name equity space, and that spanned by ϵ_t as name residual space.
We denote the portfolio weights in name equity space as $w_t^{R,\text{ name}}$ and portfolio weights in name residual space as $w_t^{\epsilon,\text{ name}}$. These weights are related by

$$w_t^{R,\text{ name}} = \Phi_t^T w_t^{\epsilon,\text{ name}} \tag{5}$$

, directly following the equality in portfolio return,

$$\left(w_t^{\epsilon\text{ name}}\right)^T \epsilon_t = \left(w_t^{\epsilon,\text{ name}}\right)^T \Phi_t \left(r_t - r_f\right) = \left(w_t^{R,\text{ name}}\right)^T \left(r_t - r_f\right) \tag{6}$$

For factors derived by PCA, we have

$$\Phi_t \beta_t = 0 \Longrightarrow \left(w_t^{R,\text{ name}}\right)^T \beta_t = \left(w_t^{\epsilon,\text{ name}}\right)^T \Phi_t \beta_t = 0, \quad \forall w_t^{\epsilon,\text{ name}} \tag{7}$$

with proof given in the appendix. It means that for any $w_t^{\epsilon,\text{name}}$, the $w_t^{R,\text{name}}$ calculated by eq. (5) satisfy,

$$\left(w_t^{R,\text{ name}}\right)^T (r_t - r_f) = (w_t^{\epsilon,\text{ name}})^T \Phi_t (\beta_t F_t + \epsilon_t) = (w_t^{\epsilon,\text{ name}})^T \Phi_t \epsilon_t = \left(w_t^{R,\text{ name}}\right)^T \epsilon_t \tag{8}$$

It suggests that the return of our statistical arbitrage portfolios is independent of market factors and relies solely on residual returns, a property usually termed as market neutrality. Ideally, portfolios are also desired to have a zero net value, known as dollar neutrality. Empirical evidence suggests that market-neutral portfolios are also approximately dollar-neutral.

We are given a factor model for stock returns:

$$r_t - r_f = \beta_t F_t + \epsilon_t, \quad t = 1, 2, \ldots, T$$

where:

- $r_t = \{r_{i,t}\}_{i=1}^N \in \mathbb{R}^N$ represents the vector of dividend-adjusted daily returns of N stocks at time t,
- $r_f \in \mathbb{R}$ is the risk-free rate,
- $F_t \in \mathbb{R}^K$ is the vector of K factors at time t,
- $\beta_t \in \mathbb{R}^{N \times K}$ is the matrix of factor loadings,
- $\epsilon_t \in \mathbb{R}^N$ represents the residual returns (unexplained component).

Our goal is to extract factors F_t statistically using PCA from the returns data, $r_t - r_f$, and to select the number K based on the eigenvalue spectrum of the empirical correlation matrix.
MODUS OPERANDI:
Step 1: Center the Returns Data
1. Compute the excess returns:

$$\tilde{r}_t = r_t - r_f, \qquad \text{for } t = 1, 2, \ldots, T$$

2. Construct the returns matrix $\mathbf{R} \in \mathbb{R}^{T \times N}$:

$$\mathbf{R} = \begin{pmatrix} \tilde{r}_1^T \\ \tilde{r}_2^T \\ \vdots \\ \tilde{r}_T^T \end{pmatrix}$$

where each row $\tilde{r}_t^T \in \mathbb{R}^N$ represents the excess returns of all stocks on day t.

Step 2: Compute the Empirical Correlation Matrix

1. Standardize $\mathbf{R}$ (if necessary) so that each column has mean 0 and standard deviation 1. Let's denote the standardized version as $\tilde{\mathbf{R}}$.

2. Compute the empirical covariance (or correlation) matrix $\mathbf{C} \in \mathbb{R}^{N \times N}$:

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

This matrix $\mathbf{C}$ captures the co-movement of the excess returns across the N stocks.

Step 3: Perform PCA on the Empirical Correlation Matrix

1. Perform an eigenvalue decomposition on $\mathbf{C}$:

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

where:

- $\mathbf{V} \in \mathbb{R}^{N \times N}$ is the matrix of eigenvectors,
- $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_N)$ is the diagonal matrix of eigenvalues, sorted such that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N$.

2. Select the Number of Factors K : Choose K based on the eigenvalue spectrum. For example, you might select K such that the cumulative proportion of variance explained by the first K eigenvalues exceeds a certain threshold (e.g., 80% or 90%).

Step 4: Construct the Factors

1. Let $\mathbf{V}_K \in \mathbb{R}^{N \times K}$ be the matrix containing the first K eigenvectors.

2. Compute the factors F_t as:

$$F_t = \mathbf{V}_K^T \tilde{r}_t, \quad \text{for } t = 1, 2, \ldots, T$$

Here, $F_t \in \mathbb{R}^K$ represents the K principal components at time t.

Step 5: Interpretation in Terms of Portfolio Weights

From Equation (2), we interpret the factors F_t as portfolios of stocks:

$$F_t = \omega_t (r_t - r_f)$$

where $\omega_t \in \mathbb{R}^{K \times N}$ contains the portfolio weights. In the PCA setup: - ω_t corresponds to $\mathbf{V}_K^T$, which gives the linear combination of stocks (or loadings) used to construct the factors.