**Energy vs. Tech: Quantifying Divergent Risk Drivers in S&P 500 Through Independent Component Analysis**

Final Project

XiLong, JinhaoGuo, HaoranLi

Humboldt-Universität zu Berlin

Digital Economy& Data Analysis

Professor WolfGang

July 20th, 2025

**Abstract**

Inthis project, we will work on the S&P 500 component stock-related financial data, which can be used for financial performance cluster analysis (the corresponding folder name implies), such as exploring different industries, different financial performance characteristics of company clusters; also can be used for statistical analysis of stock yield, risk and other dimensions, modeling predictions, such as analyzing the relationship between yield and risk (standard deviation, etc.), differences in different industry characteristics, etc.

**Field Meaning**

Company:

A company is a short name that represents a specific listed company in the S&P 500, used to distinguish between different entities.

Sector:

Industry classifications, such as Industrials, Health Care, etc., for easy analysis by industry dimension.

annual\_return\_log: Logarithmic processing of annualized returns, allowing data to be more consistent with certain statistical assumptions (such as normality convergence), reflecting the annual return of an investment - Std: standard deviation, measuring the degree of discrepancy of the data, where it reflects the risk of volatility of stock returns, the greater the value, the higher the risk.

Skewness: bias, describing the asymmetry of the data distribution, positive bias represents the right long tail, negative bias represents the left long tail, reflecting the morphological characteristics of the yield distribution.

Kurtosis: Peak, which measures the steepness of the distribution (relative to the normal distribution), with high/low peaks reflecting the concentration of data near the average or at the extreme.

FastICA (Rapid Independent Component Analysis) is used to separate independent components from mixed signals, requiring data to be a linear mix of multiple independent source signals. Here, data is a multi-dimensional characteristic of different stocks (indicators of yield, risk, etc.). If each stock is considered a "mixed signal" source, you can theoretically try to use FastICA to mine potentially independent components (e.g. yield under different drivers, risk patterns).

**OurWork**

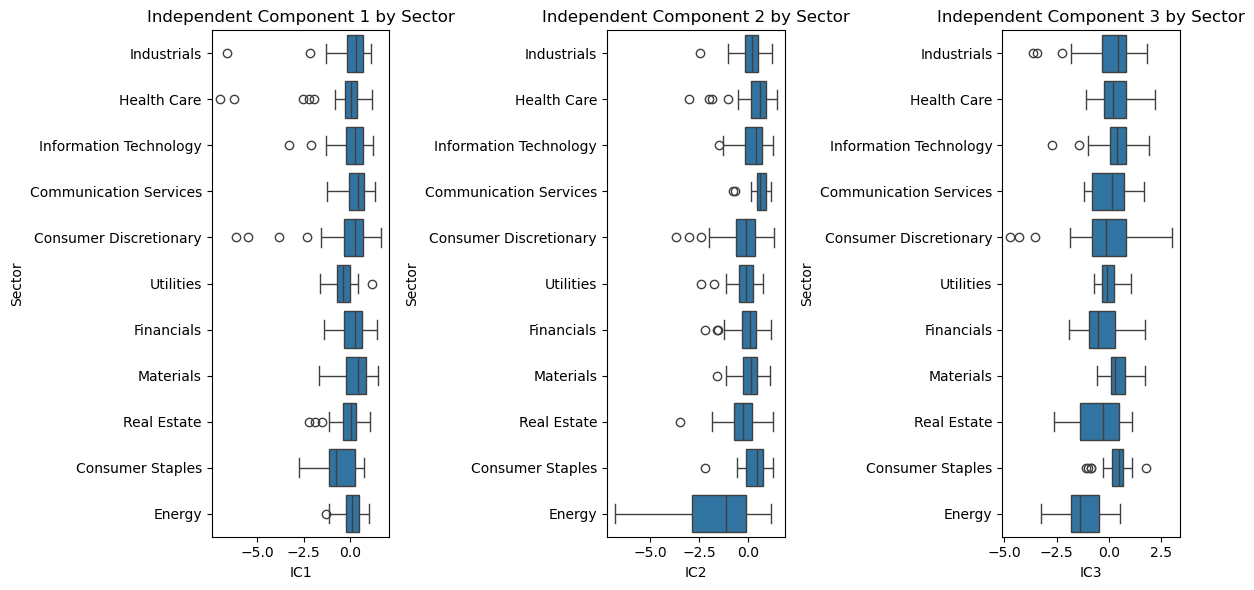
Doing analysis based on data ‘SDA\_2020\_St\_Gallen.csv’ provided by the Quartinar Webpage

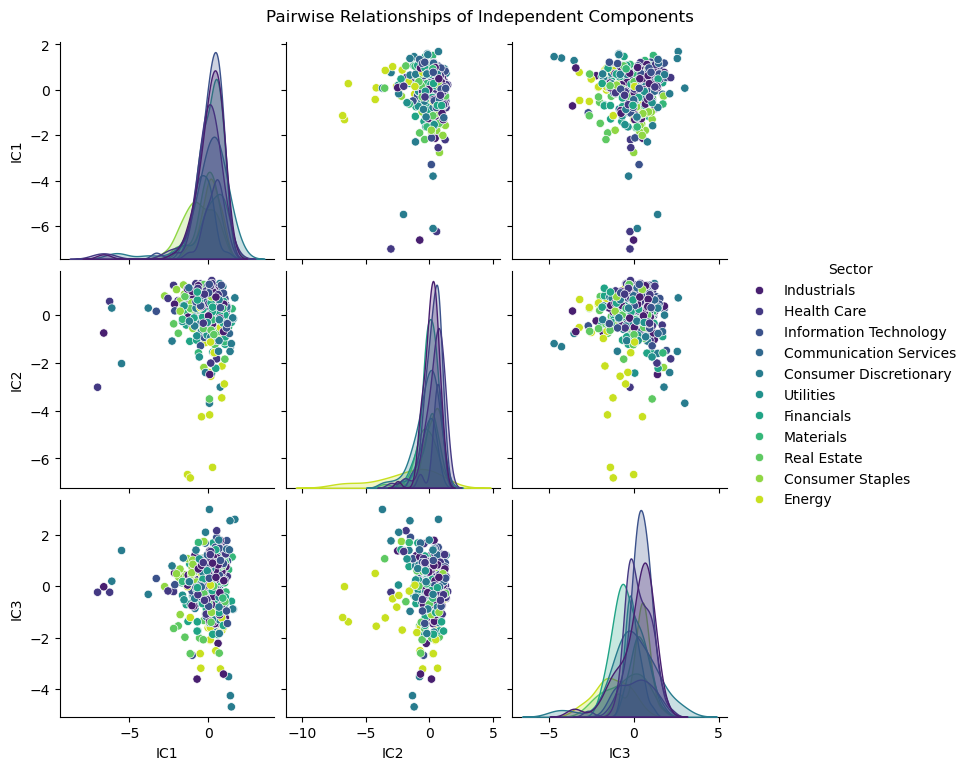
The confusion matrix is as follows:

|  |  |  |  |
| --- | --- | --- | --- |
|  | IC1 | IC2 | IC3 |
| Annual\_Rerturn\_Log | 0.030553 | 0.390005 | 0.800339 |
| Standard\_Deviation | 0.255703 | -0.495417 | -0.650653 |
| Skewness | -0.557034 | 0.780794 | -0.215736 |
| Kurtosis | -0.552071 | -0.817559 | 0.143333 |

Annual\_return\_log = log\_(Annual\_Return)

By programming, the boxplot and the Kernal density plot are shown as follows.





**ICA Analysis**

IC1 may represent “market systemic risk” and IC2 may represent “industry-specific risk patterns.”

From the plot we can see that the IC1 is very sensitive to Industrial indicator, Information indicator and Energy indicator.

IC2: Energy industry box has a low value, real estate and optional consumption is relatively dispersed(离散的), indicating that IC2 may capture the difference between the energy industry and other industries in a certain potential dimension (e.g. different drivers of the energy cycle and consumption / real estate).

IC3: Real Estate box size is wide, high discreteness, and the range of energy values is large.This reflect that IC3 describes the potential factors in the real estate and energy industries more complex, with significant differences in the industry.

In the future, we can add financial indicators such as liquidity, valuation indicators, etc. for a more comprehensive analysis of stock returns.

Industry clustering: Points in the same industry have a certain aggregation in the multi-IC space, but there are also cross-sections (e.g. data points in the info technology and healthcare industries are partially overlapping). This illustrates:

Companies in the industry behave similarly (clustering) on the potential dimensions of independent component mapping, validating the association of ICs with industry characteristics;

There are cross-sections between industries, and I speculate that certain industries are driven by common market factors (e.g. technology + healthcare are both influenced by innovation and policy).

In conclusion, these independent components likely represent the latent factors driving stock returns:

IC1 may be related to market cycles/fundamental asset volatility (outliers in energy and consumer staples correspond to cyclical industries);

IC2 may be associated with economic growth/consumer-driven factors;

IC3 may reflect risk appetite/asset allocation(the dispersion in real estate and energy corresponds to high-volatility/cyclical sectors).

In the future, companies can pay more attention on the combination of the information technology field and the healthcare field, boosting the increment of the annual stock return.

**Reference**

SDA\_2020\_St\_Gallen/sp500\_dataset.csv

https://quantinar.com