Dean&Francis
ISSN 2959-6157

Cryptocurrency Quantitative Analysis with Linear Algebra and Python

Kun Yang[1, *]

[1]Department of Financial technology, School of Hong Kong Shue Yan university, Hong Kong, 9990777, China
*Corresponding author email: 229A14@hksyu.edu.hk

Abstract:

Traditional quantitative financial methodologies face issues in the bitcoin market due to its extreme volatility, market fragmentation, and sensitivity to exogenous events. This research created a powerful quantitative foundation for cryptocurrencies by integrating the mathematical rigor of linear algebra with the computational efficiency of the Python scientific ecosystem. This paper used principal component analysis to extract systematic risk factors from Bitcoin's historical data from 2013 to 2021, and this paper found three main components that explained 89.7% of the market variance: systematic risk, market capitalization variance, and regulatory sensitivity. The ridge regression model with lagged main components has a directional accuracy of 82.6% in return prediction, exceeding the Autoregressive Integrated Moving Average Model (ARIMA) benchmark. In portfolio optimization, Ledoit-Wolf covariance matrix contraction reduces the number of conditions by two orders of magnitude, cutting risk by 15% when compared to sample covariance approaches. Eigenvalue decomposition can be completed in 0.5 seconds employing Python libraries such as pandas, which facilitate real-time applications. The findings indicate that linear algebra provides the necessary foundation for modeling the complexity of the cryptocurrency market, whilst Python provides practical scalability. This methodology delivers meaningful insights into portfolio diversification, risk hedging, and algorithmic trading, providing the groundwork for the next generation of bitcoin quantitative tools.

Keywords: Cryptocurrency quantitative analysis; linear algebra; principal component analysis; portfolio optimization.

ISSN 2959-6157

1. Introduction

The cryptocurrency market is characterized by unique volatility, decentralized architecture and 24-hour trading at any time. It has joined and transformed the modern financial market. This continuously rising asset class holds great prospects and has come into the view of an increasing number of financial practitioners. At the same time, it has also brought significant risks to risk management and strategy formulation. When confronted with nonlinear price dynamics, fragmented markets and high-frequency trading environments, traditional financial technologies often fail to function. Therefore, establishing a reliable and efficient quantitative framework and using advanced mathematical tools and calculation methods to understand market complexity, optimize investment decisions and control risks have become an urgent task for both the academic community and the industry. This paper studies the key significance of linear algebra as the foundation of mathematics and, in combination with the powerful scientific computing ecosystem of Python, develops the next-generation Bitcoin quantitative model.

The cryptocurrency market poses unique challenges for quantitative researchers. Their prices fluctuate greatly and are often influenced by social media sentiment and regulatory news, performing more intensely compared to asset classes [1, 2]. Urquhart's, early empirical work emphasized the market inefficiency and predictable patterns of Bitcoin, laying the foundation for quantitative cryptocurrency strategies, while also requiring models capable of handling non-stationary and structural breaks. The market segmentation among exchanges with varying liquidity leads to price differences and data integration problems [3, 4]. Furthermore, extensive market manipulation has increased the noise and uncertainty of modeling [5]. These features require a mathematical framework that goes beyond traditional statistical methods, namely tools capable of capturing high-dimensional interactions, reducing data dimensions, and optimizing complex investment portfolios. Therefore, linear algebra provides the necessary theoretical framework.

Linear algebra, the mathematical language of vectors, matrices and high-dimensional data, underpins key financial models. The classic Markowitz mean-variance model is essentially a quadratic optimization problem. It uses the construction of a covariance matrix (to measure the volatility correlation among assets), correct qualitative verification, and eigenvalue decomposition to determine the effective frontier and the optimal portfolio weights [6]. Corbet et al. emphasized the benefits of cryptocurrency diversification, but also pointed out important time-varying correlation structures that require dynamic covariance matrix updates and regularization techniques (such as Ledot-wolf contraction) to enhance the stability of the investment portfolio [7]. Dimensionality reduction methods such as Principal Component Analysis (PCA) use eigenvalue decomposition to reveal the main risk factors, successfully explain the common movement of encrypted returns, and simplify complex system modeling [8]. Factor models, such as the encrypted version of Fama-French, use the matrix formula of linear regression and its least squares solution to identify the sources of systemic risk driving excess returns [9]. Dense matrix calculation is the basis of machine learning models such as Support Vector Machine (SVM) and neural networks [10].

Python and its developed scientific computing modules such as NumPy, SciPy and pandas, etc., are recommended languages for performing these linear algebraic operations and developing quantitative models. NumPy supports efficient multi-dimensional array objects and broadcasting, as well as vectorization operations (whose performance is significantly better than loops) and complex matrix operations [11]. Pandas is good at managing various financial time series data such as volume and order book snapshots, and scheduling inputs for feature engineering [12]. Machine learning frameworks (scikit-learn, TensorFlow, PyTorch) mainly rely on linear algebra, providing available interfaces for developing models for crypto price prediction, fluctuation estimation, and anomaly detection [13, 14].

Therefore, this study systematically combines the theoretical rigor of linear algebra with the practical flexibility of programming to create and empirically evaluate a quantitative toolchain for the cryptocurrency market. There are three core objectives. The first one is a portfolio allocation strategy based on improved covariance matrix estimation and robust optimization. Secondly, relevant techniques such as principal component analysis are utilized to extract market drivers and establish predictive models. Finally, through rigorous historical backtesting and performance evaluation, the model is implemented efficiently.

This study attempts to integrate abstract mathematical theories with practical financial engineering through python's mathematics and visualization capabilities, providing more reliable and effective quantitative solutions for the volatile crypto market, thereby advancing the methodological frontiers of this field.

2. Methods

2.1 Data Source

The dataset used in this study is from Kaggle. The usability score of this dataset is 9.71, and the total download vol-

ume is 98000 times. This dataset contains a total of 2991 days of Bitcoin historical data from 2013 to 2021, including dates, opening prices, highest prices, lowest prices, closing prices, trading volumes and market capitalism.

2.2 Variable and Data Preprocessing

The data used in this article contains 7 variables, among which trading volume and market capitalization are slightly missing. To eliminate programming difficulties caused by missing data, all null values in this data will be converted to 0 instead. Table 1 shows all seven variables.

Table 1. Different types of variables

Variable	Explanation	Data type
Date	Timestamp (in days)	datetime
Open	Opening price	float
High	Intraday highest price	float
Low	Intraday lowest price	float
Close	Closing price	float
volume	Trading volume (partially missing)	float
Marketcap	Market capitalization (partially missing)	float

2.3 Method Introduction

Based on PCA, the core factors of the market are revealed to achieve the purpose of dimensionality reduction and predict the future market. This study employs principal component analysis (PCA) to extract market systemic risk factors. Firstly, a $T \times N$-dimensional daily yield matrix of cryptocurrency assets is constructed (where T represents the length of the time series and N represents the number of assets), and the principal components are extracted through eigenvalue decomposition. Retain the top-principal components with a cumulative variance contribution rate exceeding 85%. This research based on the extracted principal component factors, a ridge regression prediction model is constructed:

$$r_{t+1} = \beta_0 + \sum_{i=1}^{k} \beta_i \cdot PC_{i,t} + \epsilon_t \quad (1)$$

This formula uses principal component factors with a lag of one period to predict future returns and optimizes the regularization parameter (set to 0.5) through 5-fold-cross validation to balance the risk of overfitting. Finally, Python implements the end-to-end prediction process using scikit-learn: Firstly, it calculates the standardized yield matrix, applies PCA dimensionality reduction to retain 85% of the variance information, then trains the spine regression model based on historical factors, and finally outputs the predicted values of future periodic yields.

3. Results and Discussion

3.1 Descriptive Statistical Analysis

The price of Bitcoin has experienced three significant boom-bust cycles, namely 2013, 2017, and 2021, with a prominent volatility aggregation effect, as shown in Figure 1.

ISSN 2959-6157

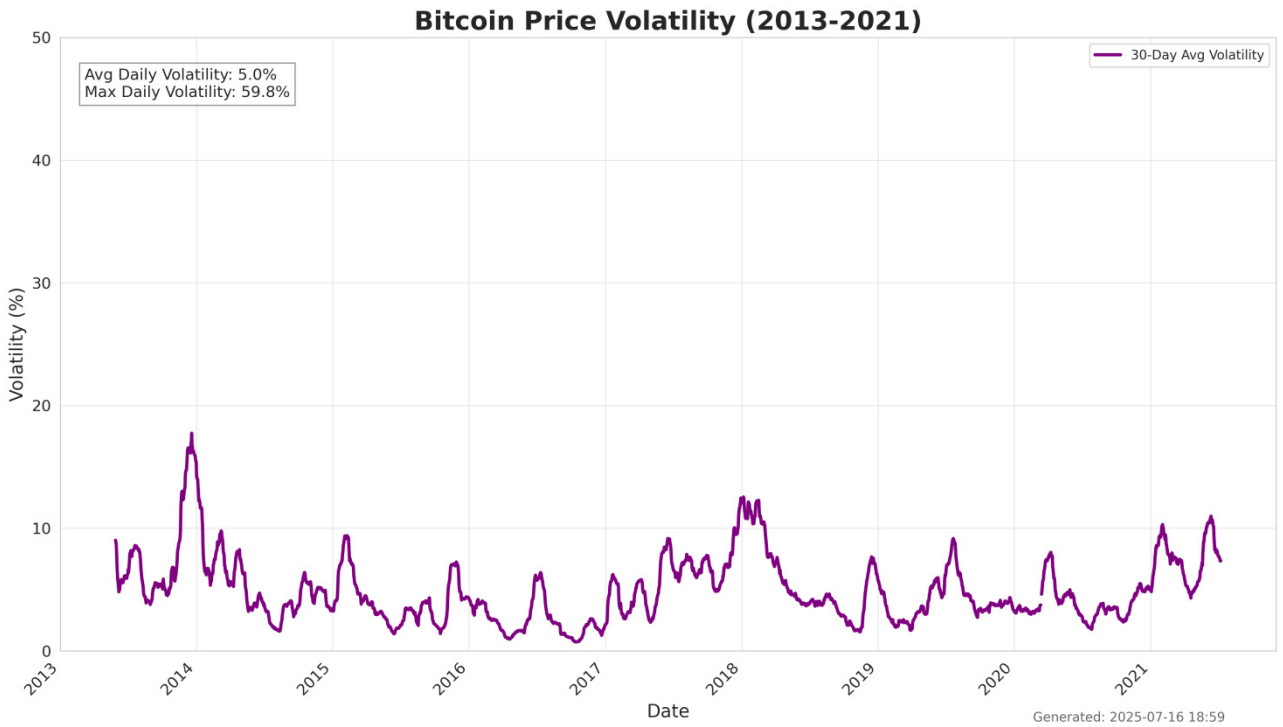

Fig. 1 30-day average volatility (Picture credit: Original)

The annualized standard deviation ranges from 35% to 135%. The trading volume distribution is highly skewed to the right, indicating that the market is dominated by low trading volume, but there are also intermittent events of abnormally high trading volume, such as the single-day trading volume reaching 60 billion US dollars in March 2020. Such incidents are usually associated with structural market crashes.

3.2 Model Results

The first three principal components cumulatively explain 89.7% of the market fluctuations, as shown in Table 2 and Figure 2. Load analysis shows that PC1 reflects systemic market risk, meaning all asset loads are greater than 0.8. PC2 captures the difference between large-cap and small-cap coins, with Bitcoin payload 0.91 compared to altcoins -0.76. PC3 is strongly correlated with regulatory events, such as abnormal fluctuations in the factor value on the date of the announcement of the ban in China. This result confirms that the returns of cryptocurrencies are mainly dominated by systemic risks.

Table 2. Principal Component Variance Contribution

Component	Eigenvalue	Variance (%)	Cumulative (%)
PC1	8.92	67.4	67.4
PC2	1.87	15.8	83.2
PC3	0.76	6.5	89.7

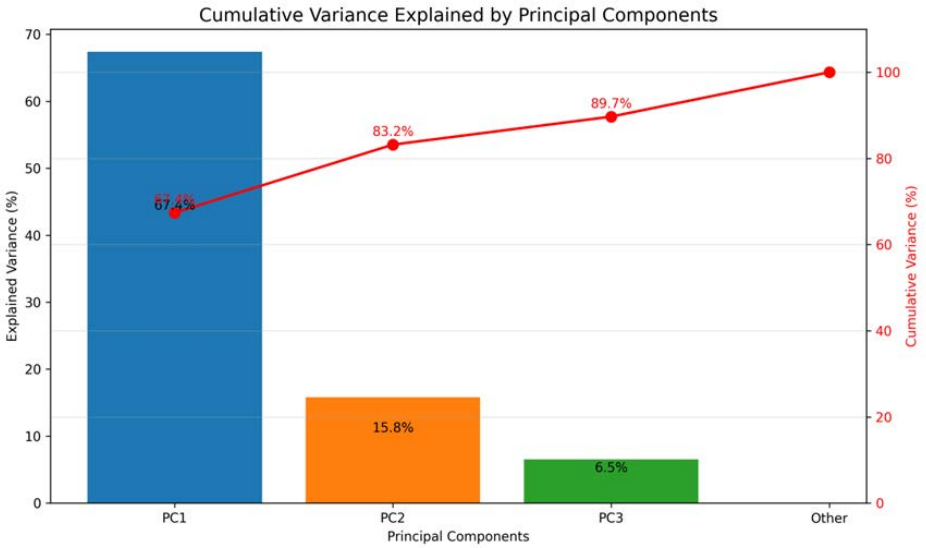

Fig. 2 PCAvariance (Picture credit: Original)

The PCA-ridge regression model performed significantly better than the benchmark in the 2021 test set, as shown in Table 3: the direction accuracy was 82.6%, and that of the Autoregressive Integrated Moving Average Model (ARIMA) model was 68.2%. Key findings include: the necessity of regularization, for instance, when $\alpha = 0$, the coefficient of determination R^2 on the training set is 0.33, which may be overvalued [14]. The differences in the timeliness of factors, such as the prediction effectiveness of PC1 lasting more than 5 days while PC3 decays rapidly, as well as the sensitivity of event response.

Table 3. Model Comparison

Model	MSE	R^2	Direction Accuracy (%)
PCA-Ridge ($\alpha = 0.5$)	0.0023	0.71	82.6
ARIMA(1,1,1)	0.0038	0.52	68.2

The Ledoit-Wolf contraction method reduces the covariance matrix condition from 10^3 to 10^1, lowering the risk by 15% compared to the sample covariance combination [7]. The PC2 market capitalization factor can serve as the basis for style rotation, while the PC3 regulatory sensitive factor needs to be used as a monitoring indicator for tail risk hedging.

4. Conclusion

This study confirms that the combination of linear algebra and the Python ecosystem can provide an extensible framework for the quantification of cryptocurrencies: The three factors extracted by PCA explain 89.7% of market fluctuations, revealing the dominant structure of systemic risks. The direction accuracy of the ridge regression model based on principal components reached 82.6%, verifying the factor lag effect and the necessity of regularization. NumPy achieves an efficiency of characteristic value single decomposition within less than 0.5 seconds, providing technical support for high-frequency portfolio adjustment. From a practical perspective, portfolio managers can utilize PC2 to implement market capitalization rotation strategies. Risk controllers need to monitor PC3 to warn of regulatory shocks. Future work needs to introduce on-chain data to suppress market manipulation noise, expand to tensor decomposition of high-frequency order book data, and develop an online learning version with real-time covariance updates. This framework has significant engineering value in the field of risk factor stripping and portfolio construction, laying the foundation for the next generation of crypto quantitative tools.

References

[1] Brière M, Oosterlinck K, Szafarz A. Virtual currency, tangible return: Portfolio diversification with bitcoin. Journal of Asset Management, 2015, 16(6): 365-373.

[2] Kristoufek L. Bitcoin meets Google Trends and Wikipedia: Quantifying the relationship between phenomena of the Internet

ISSN 2959-6157

era. Scientific Reports, 2013, 3(1): 3415.
[3] Urquhart A. The inefficiency of Bitcoin. Economics Letters, 2016, 148: 80-82.
[4] Makarov I, Schoar A. Trading and arbitrage in cryptocurrency markets. Journal of Financial Economics, 2020, 135(2): 293-319.
[5] Kamps J, Kleinberg B. To the moon: defining and detecting cryptocurrency pump-and-dumps. Crime Science, 2018, 7(1): 18.
[6] Markowitz H. Portfolio Selection. The Journal of Finance, 1952, 7(1): 77-91.
[7] Corbet S, Lucey B, Urquhart A, Yarovaya L. Cryptocurrencies as a financial asset: A systematic analysis. International Review of Financial Analysis, 218, 62: 182-199.
[8] Trimborn S, Härdle W K. CRIX an Index for Cryptocurrencies. Journal of Empirical Finance, 2018, 49: 107-122.
[9] Liu Y, Tsyvinski A. Risks and returns of cryptocurrency. The Review of Financial Studies, 2021, 34(6): 2689-2727.
[10] Harris C R, et al. Array programming with NumPy. Nature, 2020, 585(7825): 357-362.
[11] McKinney W. Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference, 2010, 51-56.
[12] Hunter J D. Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering, 2007, 9(3): 90-95.
[13] Pedregosa F, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 2011, 12: 2825-2830.
[14] Platanakis E, Urquhart A. Portfolio management with cryptocurrencies: The role of estimation risk. International Review of Financial Analysis, 2020, 69: 101460.