Bachelor’s Thesis Proposal

Institute of Economic Studies
Faculty of Social Sciences
Charles University in Prague

Author’s name and surname: Tomáš Barhoň

E-mail: tomas.barhon@hotmail.cz

Phone: +420 724 770 037

Supervisor’s name: prof. PhDr. Ladislav Krištoufek Ph.D.

Supervisor’s email: ladislav.kristoufek@fsv.cuni.cz

Proposed Topic:

Noise reduction and feature extraction with principal component analysis for cryptocurrency price modeling

Preliminary scope of work:

Research question and motivation

Crypto assets have always been exceptionally volatile compared to traditional assets such as stocks or gold. The historical window is relatively short, thus modeling their price or volatility proposes quite a difficult challenge. It is generally believed that noise in any data decreases the precision of predictions. This effect might be reduced, which will improve the performance of traditional models that are used for cryptocurrency price modeling.

The main motivation for researching this topic is that there is still an ongoing discussion about the role of different features in crypto pricing dynamics. (Kukacka; Kristoufek 2023) have shown that a lot of the pricing dynamic emerges from complex interactions between fundamental and speculative components. They also show the different correlations between all of the explanatory variables which have a direct connection to principal component analysis. It is crucial to study the real impact of those variables in different models as many of them might turn out to be obsolete.

There is currently little use of this dimensionality reduction technique in the academic literature about cryptocurrencies. However, for more traditional financial series this technique is already quite established as a preprocessing technique to reduce noise and dimensionality from which financial data inherently suffer (Chowdhury, U. ; Chakravarty, S. and Hossain, M. 2018). Moreover (Bouri, E.; Kristoufek, L.; Ahmad, T. et al. 2022) studied the effect of microstructural noise on idiosyncratic volatility in cryptocurrencies which further supports the need for a technique that will mitigate this effect on the predictions.

The research will address the problem of variable selection for different types of predictive models with respect to the analysis of the principle components aiming to reduce the dimensionality and simultaneously increase precision. The second question is whether it is more appropriate to transform the high dimensionality with PCA into lower dimensionality or simply omit the variables with high multicollinearity from the models. These approaches are fundamentally different and the answer is not clear.

Contribution

Existing research agrees that financial data and especially cryptocurrency data are significantly affected by noise. The main goal is to extend the research on the topic of variable selection for algorithmic trading models as there are still a lot of unanswered questions. It will most likely become clearer which approach to dimensionality reduction is the most efficient concerning cryptocurrencies.

Also, not only the underlying pricing dynamics will be detected but the results can be used for investors that are trying to lower their risk of loss which is relatively high in crypto markets. The effect of having a more stable and precise model might significantly cut the transaction costs that are associated with more frequent exchanges as the predictions will become less volatile. That is a desirable property needed to maximize profit and increase credibility towards its customers.

Methodology

The data will come from various sources because the aim is to look at all the possible variables even if they might not seem useful at first glance. As already mentioned the dynamic is driven by a lot of completely different effects. Most of it will be collected from: coinmetrics.io, studio.glassnode.com, and for macroeconomic indicators https://fred.stlouisfed.org/. Some of the data might need to be interpolated to daily observations. Lastly, the observations will need to be sliced to different window sizes and shifted by one so that the predictions can be made for the next day with the data available on that day.

Afterward, the multicollinearity in the data will be examined and different approaches to solve it will be used. The two main ones are using only a smaller set of uncorrelated variables (simple dimensionality reduction) and the second being employing PCA transformation to preserve a predefined threshold of variance or directly targeting the number of principal components.

All the different setups will be compared across three models: linear regression, SVM, and LSTM neural network. The hypothesis is that PCA transformation will substantially lower the measured errors for linear regression and SVM although for LSTM it will result in lower performance as it will only decrease the capacity of the model as the model is powerful enough to create such uncorrelated features without the PCA transformation.

Outline

Abstract
Introduction
1. why should we study this topic
2. is there any existing knowledge about the topic
3. the contribution to existing research
4. main results and outcomes of the thesis
5. what will be the following structure of the text

Data
1. how was the data collected (period, interpolation, different sources)
2. which transformations were applied
3. different windows (25,50,100 days)

Literature review and hypotheses
1. literature on PCA in time series generally
2. literature on the pricing dynamics of cryptocurrencies (for variable selection)
3. literature on different uses of PCA in a crypto context
4. formulation of research hypotheses

Methodology
1. how will the problem be analyzed from different perspectives (more generally the logic of why it should work)
2. presentation of 3 baselines models (from simple to more complex) and their performance with all the window sizes
3. comparison of the approaches of using PCA on all variables vs. not using variables that suffer from multicollinearity on all of the models after the data transformations

Results
1. whether to reject the hypotheses
2. interpretation of the result

Conclusion
1. conclusion of the main findings
2. whether there is a use for the results found in practice
3. opportunities for future research

Bibliography

1. Kukacka, J., & Kristoufek, L. (2023). Fundamental and speculative components of the cryptocurrency pricing dynamics (Vol. 9). Financial Innovation.
2. Kristjanpoller, W., & Minutolo, M. C. (2018). A hybrid volatility forecasting framework integrating GARCH, artificial neural network, technical analysis and principal components analysis (Vol. 109). Expert Systems with Applications.
3. Chowdhury, U. N., Chakravarty, S. K., & Hossain, M. T. (2018). Short-Term Financial Time Series Forecasting Integrating Principal Component Analysis and Independent Component Analysis with Support Vector Regression (Vol. 6).
4. Bouri, E., Kristoufek, L., , & Shahzad, S. J. H. (2022). Microstructure noise and idiosyncratic volatility anomalies in cryptocurrencies. Springer Link. https://doi.org/10.1007/s10479-022-04568-9
5. Rea, A., & Rea, W. (2016). How many components should be retained from a multivariate time series PCA?. *arXiv preprint arXiv:1610.03588*.