WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

Master Thesis

Edit: Master Thesis Title

Jorge Eduardo Frías Navarrete

Submitted in partial fulfillment of the requirement for the degree of:

Master of Science

Student ID:	012329686
Degree programme:	Quantitative Finance
Supervisor:	Univ.Prof. David Preinerstorfer, Ph.D.
Date of Submission:	August 18, 2025

Department of Finance, Accounting and Statistics.
Vienna University of Economics and Business.
Welthandelsplatz 1, 1020 Vienna, Austria.

Acknowledgements

Here write acknowledgements.

Abstract

Here goes my abstract text. Here goes my abstract text. Here goes my abstract text. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Suspendisse eu dolor luctus, rhoncus leo in, commodo turpis. Aenean sed enim in sem euismod porta. Vivamus tempor lorem nec eros rhoncus, eu hendrerit libero tincidunt. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Quisque dapibus turpis quis nibh molestie dapibus. Aliquam erat volutpat. Integer et odio nec mauris sollicitudin mattis.

Table of contents

List of figures

List of tables

1. Introduction

Mirar Liu et al (y el otro paper similar) donde habla de las criptomonedas, e inspirarnos de ahi. Quiza la otra tesis en esto tambien.

Empezar con una introduccion de criptomonedas, del mercado, de la gran volatilidad, grandes retornos. Mencionar coin Market cap, la capitalizacion del mercado total de criptomonedas.

Mencionar articulos o reportes donde mencionen la importancia de este mercado, cuantas personas en promedio tienen criptos en su portafolio. Mencionar sucesos recientes importantes, como la introduccion de criptomonedas en algunos exchanges, de futuros en CME, de indices en XXX, del boom en la crisis de COVID (ver Mercik, donde menciona sucesos importantes).

mencionar a Ross, que comprobo la estructura lineal de los factores: – Esto quiza en introduccion

** Ejemplo de frase See (**?**) for additional discussion of literate programming.**

```
1 + 1
```

```
[1] 2
```

2. Methodology

The Instrumented Principal Component Analysis (IPCA) model was introduced in the seminal work of Kelly et al. (2019, 2020). The main model used in this thesis is the IPCA with different K number of factors

Explicar la metodologia de IPCA. Si hay tiempo, entonces explicar tambien como funciona el RPCA de Chen and Roussanov. Explicar las R2 (en lugra de R2, entonces poner Total score y predictive score), mencionar como pie de pagina que son las medidas definidas por Kelly et al. (2019).

Explicar los bootstrap para medir la significancia cada caracteristica, y quiza mencionar tambi'en brevemente los characteristic managed portfolios, en que consisten y como se emplean (quiza tomar inspiracion de Kelly, Bianchi, o creo que puede ser mejor en Liu et al.)

2.1. Instrumented Principal Component Analysis

3. Data

In this section, I introduce the cryptocurrency data used in this thesis, the series of filters applied to clean and prepare the dataset, and the summary statistics of the cryptocurrency excess returns. In addition, I show the set of asset-specific characteristics constructed from the cryptocurrency market data, which are used as instruments for latent factor exposures in the IPCA model. Appendix A.2 provides a detailed description of the characteristics used in the empirical analysis.

[++++ **ADD SMALL INTRO ABOVE OF RIK FACTORS CREATED** ++++]

The data extraction and pre-processing are primarily conducted in R 4.5.1 (R Core Team, 2025), using, among other packages[1], the `tidyverse` (v. 2.0.0; Wickham et al., 2019). Additional cleaning steps and visualizations are performed in Python 3.13.5 (Python Software Foundation, 2025). The full reproducible code is available in Appendix A.1.

3.1. Data extraction and sample construction

I collect daily cryptocurrency data on open, high, close, and low (OHCL) prices, 24-hour volume, and market capitalization (calculated as the cryptocurrency's USD price multiplied by its circulating supply) from CoinCodex, a website-data provider that gathers and aggregates data from more than 400 exchanges. I extract the data, all expressed in US dollars, using the CoinCodex API as follows:

1. I retrieve the list of all available cryptocurrencies and extract each cryptocurrency shortname, also referred to as the "slug". At the time of writing, there are 14,907 unique cryptocurrency shortnames listed in the API.

2. Using the slug, I construct an URL for each cryptocurrency to obtain the metadata from the API. I parse the JSON API response into a dataframe and extract

[1] See Appendix A.4 for the full list of software used in the empirical study.

the OHCL prices, volume, and market capitalization daily data. I exclude those observations with non-zero or missing values in any of these fields.

Out of the 14,907 cryptocurrencies listed, only 7,272 entries contained available data. Next, following the methodology of Bianchi & Babiak (2021) and Mercik et al. (2025), I apply a series of cleaning and filtering steps in order to remove possible innacuracies in the dataset:

1. Non-positive and missing values. As mentioned earlier, I remove observations where prices, volume, or market capitalization were non-positive or missing.

2. Small cryptocurrencies. Similar to Liu et al. (2022), I screen out small cryptocurrencies and consider only those with a market capitalization greater than one million USD. Therefore, I exclude observations for coins whose market capitalization falls below this minimum threshold, which allows for the possibility that a coin may become "small" after a certain period or event.

3. Cryptocurrency type. Based on the cryptocurrency classification from CoinMarketCap and CoinCodex, I exclude:

    - stablecoins. I include (i) centralized stablecoins, which are backed and pegged to fiat currency or physical assets by a third party, such as Tether (USDT), USD Coin (USDC), and Euro Coin (EURC), and (ii) algorithmically stabilized stablecoins, which use algorithms to adjust the circulating supply in response to changes in demand to maintain a stable value with the underlying asset, such as DAI and AMPL (FSB, 2020).

    - wrapped cryptocurrency tokens, which mirror the value of another cryptocurrency from a different blockchain, e.g., Wrapped Bitcoin (wBTC) or Wrapped Ethereum (wETH) (Coinbase, n.d.).

    - cryptocurrencies backed by or pegged to gold or precious metals, including Pax Gold (PAXG) or XAGx Silver Token (XAGX).

4. Erroneous trading volume. To filter out cryptocurrencies with "fake" or "erroneous" trading volume, I calculate the daily volume-to-market-capitalization ratio for each token and exclude observations where the ratio exceeds 1.

5. Extreme returns. To minimize the influence of extreme values in my results, I winsorize daily cryptocurrency returns to lie within the range of -90% to 500%.

6. Time period. Even though cryptocurrency data are available since 2014, I use data from June 1, 2018 for the empirical analysis due to the low amount of coins

available before this date (see Figure 3.1).

7. Minimum observations. In order to maintain practical relevance, I keep cryptocurrencies that have at least 365 consecutive daily observations and those with at least 730 observations in the complete panel of coin characteristics (see Section 3.3), which is equivalent to 2 years of historical data. Therefore, I exclude very short-lived coins, but retain failed coins with this relatively large number of observations, which help to lessen the so called "survivorship biais".

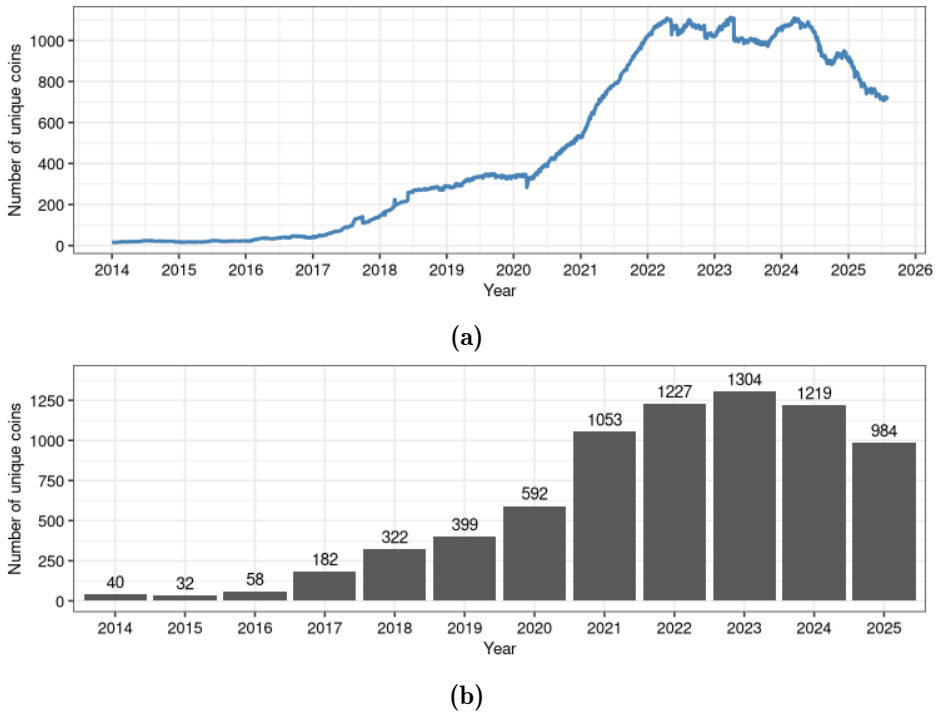

Figure 3.1.: Number of cryptocurrencies over time. Panel A shows the daily time series of the number of unique cryptocurrencies. Panel B displays the number of unique cryptocurrencies recorded each year. Both panels correspond to the dataset after applying the filtering steps (1) to (5), covering the period from January 1, 2014, to July 31, 2025, and including 1,416 unique cryptocurrencies. Note that coins may enter or exit the market over time.

3.2. Sample overview

After applying all the filters, the resulting sample consists of 973 unique cryptocurrencies and 1,478,936 observations from June 1, 2018, to July 31, 2025, where a day starts at 00:00:00 UTC. It is important to mention that the number of cryptocurrencies fluctuates over the entire period, which results in an unbalanced panel of data.

Table 3.1.: Summary statistics of daily excess returns. The table reports summary statistics of daily excess returns for the filtered sample, the top 100 cryptocurrencies ranked by market capitalization, and for Bitcoin, Ethereum, and Ripple individually. Reported statistics include the number of daily observations, the number of unique coins over the sample period, the mean and standard deviation of returns, and the 10th percentile, lower quartile, median, upper quartile, and 90th percentile of the distribution of the returns. The sample period is from June 1, 2018, to July 31, 2025.

	No. Obs	Unique coins	Mean	Std	P10	P25	P50	P75	P90
Sample	1,478,936	973	-2.70%	12.49%	-10.18%	-6.65%	-3.70%	0.02%	4.69%
Top 100	176,400	100	-2.94%	7.28%	-9.20%	-6.32%	-3.56%	-0.16%	3.67%
Top 10[1]	24,747	10	-2.35%	6.14%	-7.80%	-5.44%	-2.87%	0.22%	3.55%
Bitcoin	2,618		-2.40%	3.92%	-6.60%	-4.96%	-2.58%	-0.20%	2.26%
Ethereum	2,611		-2.40%	4.85%	-7.46%	-5.27%	-2.83%	0.27%	3.48%
Ripple	2,540		-2.41%	5.70%	-7.65%	-5.43%	-2.77%	-0.00%	2.87%

[1] As of July 31, 2025, the top 10 cryptocurrencies are Bitcoin, Ethereum, Ripple, Binance Coin, Solana, Dogecoin, Tron, Cardano, Stellar, and Chainlink.

3.3. Characteristic construction and description

Following Kelly et al. (2019), I cross-sectionally transform the instrument variables period-by-period in the following manner: first,

This is more related to factor construction.

Organize week in the following way: the first seven days of the year forms the first week, and the first 51 weeks of the year consists of 7 days each. The 52th week of the year consists of the last eight days and, in case of a leap year (as 2016, 2020, and 2024), of nine days.

Following Liu et al. (2022), I construct a daily cryptocurrency market return as the value-weighted average return of all the cryptocurrencies in the sample. For cryptocurrencies $i = 1, ..., N$, the daily market return at time t is computed as:

$$r_t^M = \frac{\sum_{i=1}^{N} r_{it} \cdot marketcap_{it}}{\sum_{i=1}^{N} marketcap_{it}}$$

The cryptocurrency market excess return is constructed as the difference between the cryptocurrency market return and the risk-free rate. To proxy the risk-free rate, I used the (daily) 1-month Treasury bill rate from the FRED.

Write this in the following section of "Empirical application" or This is for the model: 7. (Still undecisive) Minimum cross-section. Following the criterion by Kelly, I Convert variables in the -0.5 - 0.5 range

The sample period ranges from January 1st, 2014, to May 31st, 2025.

4. Results

Implemented in python, based on the IPCA python code of Seth Pruitt [1] and the `ipca` python package of Buechner & Bybee (2019) [2].

Important: mention the shift of characteristics: the conditional APT of Kelly, Pruitt, Su (JFE 2019) says that the characteristics known at Date=d-1 determine the exposures associated with the returns realized at Date=d; hence, here we should have shifted the characteristics in Z relative to the returns in R

This is a template of the table of the results of the IPCA model. I need to add a caption to the table. Here I reference Table 4.1.

Some test for quarto and latex

Quarto: 1. Sees the caption line after the table. 2. Wraps the `tabular` inside a LaTeX `table` environment. 3. Adds `\caption{Some letters with LaTeX}` and `\label{tbl-letters}` automatically. 4. Gives it a table number and puts it in the List of Tables.

Inline LaTeX way inside Quarto
Here we see the summary statistics in Table 4.2.

[1] See https://sethpruitt.net/research/.
[2] See https://bkelly-lab.github.io/ipca/.

Table 4.1.: Results of IPCA regression. Model Performance. Panel A and B report total and predictive R^2 in percent for the restricted ($\Gamma_\alpha = 0$) and unrestricted ($\Gamma_\alpha \neq 0$) IPCA model for K number of factors on daily and weekly data, respectively. Panel C reports the corresponding total and predictive R^2 for a simple PCA model on weekly data.

		K		
		$K = 3$	$K = 5$	$K = 8$
Panel C: PCA on weekly data				
R^2_{Total}		0.0000	0.0000	0.0000
$R^2_{\text{Predictive}}$		0.0000	0.0000	0.0000
Panel B: IPCA on weekly data				
R^2_{Total}	$\Gamma_\alpha = 0$	0.2625	0.2817	0.2934
	$\Gamma_\alpha \neq 0$	0.2661	0.2826	0.2937
$R^2_{\text{Predictive}}$	$\Gamma_\alpha = 0$	0.1725	0.1551	0.1511
	$\Gamma_\alpha \neq 0$	0.1719	0.1584	0.1554
Panel A: IPCA on daily data				
R^2_{Total}	$\Gamma_\alpha = 0$	0.2301	0.2509	0.2681
	$\Gamma_\alpha \neq 0$	0.2322	0.2524	0.2690
$R^2_{\text{Predictive}}$	$\Gamma_\alpha = 0$	-0.3904	-0.4082	-0.4169
	$\Gamma_\alpha \neq 0$	-0.3857	-0.4055	-0.4156

Table 4.2.: Some letters with LaTeX

A	B	C
D	E	F

5. Conclusion

References

Allaire, J. J., Teague, C., Scheidegger, C., Xie, Y., Dervieux, C., & Woodhull, G. (2025). *Quarto* (Version 1.7) [Computer software]. https://doi.org/10.5281/zenodo.5960048

Ardia, D., Guidotti, E., & Kroencke, T. A. (2024). Efficient estimation of bid–ask spreads from open, high, low, and close prices. *Journal of Financial Economics*, *161*, 103916. https://doi.org/10.1016/j.jfineco.2024.103916

Bianchi, D., & Babiak, M. (2021). *Mispricing and Risk Compensation in Cryptocurrency Returns* (SSRN Scholarly Paper 3935934). Social Science Research Network. https://doi.org/10.2139/ssrn.3935934

Buechner, M., & Bybee, L. (2019). *ipca: Instrumented principal components analysis* [Computer software]. https://github.com/bkelly-lab/ipca

Coinbase. (n.d.). *What is wrapped crypto?* Retrieved August 6, 2025, from https://www.coinbase.com/learn/your-crypto/what-is-wrapped-crypto

Financial Stability Board. (2020). *Addressing the regulatory, supervisory and oversight challenges raised by "global stablecoin" arrangements: Consultative document*. https://www.fsb.org/2020/04/addressing-the-regulatory-supervisory-and-oversight-challenges-raised-by-global-stablecoin-arrangements-consultative-document/

Harris, C. R., Millman, K. J., Walt, S. J. van der, Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., Kerkwijk, M. H. van, Brett, M., Haldane, A., Río, J. F. del, Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Kelly, B. T., Pruitt, S., & Su, Y. (2019). Characteristics are covariances: A unified model of risk and return. *Journal of Financial Economics*, *134*(3), 501–524. https://doi.org/10.1016/j.jfineco.2019.05.001

Kelly, B. T., Pruitt, S., & Su, Y. (2020). *Instrumented Principal Component Analysis* (SSRN Scholarly Paper 2983919). Social Science Research Network. https://doi.

org/10.2139/ssrn.2983919

Komsta, L., & Novomestky, F. (2022). *moments: Moments, cumulants, skewness, kurtosis and related tests*. https://doi.org/10.32614/CRAN.package.moments

Liu, Y., Tsyvinski, A., & Wu, X. (2022). Common Risk Factors in Cryptocurrency. *The Journal of Finance*, *77*(2), 1133–1177. https://doi.org/10.1111/jofi.13119

Mercik, A., Bdowska-Sójka, B., Karim, S., & Zaremba, A. (2025). Cross-sectional interactions in cryptocurrency returns. *International Review of Financial Analysis*, *97*, 103809. https://doi.org/10.1016/j.irfa.2024.103809

Moskowitz, T. J., Ooi, Y. H., & Pedersen, L. H. (2012). Time series momentum. *Journal of Financial Economics*, *104*(2), 228–250. https://doi.org/10.1016/j.jfineco.2011.11.003

Peterson, B. G., & Carl, P. (2024). *PerformanceAnalytics: Econometric tools for performance and risk analysis*. https://doi.org/10.32614/CRAN.package.PerformanceAnalytics

Posit team. (2025). *RStudio: Integrated development environment for r*. Posit Software, PBC. http://www.posit.co/

Python Software Foundation. (2025). *Python programming language* (Version 3.13.5) [Computer software]. https://www.python.org/

R Core Team. (2025). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Ryan, J. A., & Ulrich, J. M. (2025). *quantmod: Quantitative financial modelling framework*. https://doi.org/10.32614/CRAN.package.quantmod

Stacklies, W., Redestig, H., Scholz, M., Walther, D., & Selbig, J. (2007). pcaMethods – a bioconductor package providing PCA methods for incomplete data. *Bioinformatics*, *23*, 1164–1167.

The pandas development team. (2020). *Pandas-dev/pandas: pandas* [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.3509134

Vaughan, D. (2024). *slider: Sliding window functions*. https://doi.org/10.32614/CRAN.package.slider

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., ... SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller,

E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., ... Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Zeileis, A., & Grothendieck, G. (2005). zoo: S3 infrastructure for regular and irregular time series. *Journal of Statistical Software*, *14*(6), 1–27. https://doi.org/10.18637/jss.v014.i06

A. Appendix

A.1. Supplementary Material

A.2. Cryptocurrency Characteristics

A.2.1. Volume shock

Following Bianchi et al. (2022), the volume shock is defined as the log-deviation of trading volume from its rolling average (over 30 or 60 days) for cryptocurrency i at time t. For $m \in \{30, 60\}$ periods, the volume shock is estimated as:

$$v_{i,t} = \log(\text{Volume}_{i,t}) - \log\left(\frac{1}{m}\sum_{s=1}^{m}\text{Volume}_{i,t-s}\right)$$

A.3. Risk

A.3.1. Realized volatility (rvol)

Using the volatility estimator of Yang and Zhang (2000), I compute the daily realized volatility based on OHCL prices over a rolling 30-day window. For $n > 1$ number of periods, the volatility estimate at time t is:

$$\sigma_t = \sqrt{\sigma_O^2 + k\sigma_C^2 + (1-k)\sigma_{RS}^2}$$

where σ_{RS}^2 is the variance estimator of Rogers et al. (1994), and σ_O^2, σ_C^2, k are defined as follows:

$$\sigma_O^2 = \frac{1}{n-1}\sum_{i=1}^{n}(o_i - \bar{o})^2,$$

$$\sigma_C^2 = \frac{1}{n-1}\sum_{i=1}^{n}(c_i - \bar{c})^2,$$

$$k = \frac{\alpha - 1}{\alpha + \frac{n+1}{n-1}}$$

with $o = ln\, O_t - ln\, C_{t-1}$, and $c = ln\, C_t - ln\, O_t$. Here, C_{t-1} denotes the last days' closing price and O_t the current day's opening price. I set the constant $\alpha = 1.34$ as suggested by Yang and Zhang (2000) to be the best value in practice.

Moskowitz et al. (2012)

A.4. Software

This thesis was fully written using Quarto (Allaire et al., 2025), running in RStudio (v. 2025.5.1.513; Posit team, 2025) on Fedora Linux 42 (Workstation Edition).

I used R 4.5.1 (R Core Team, 2025) and the following R packages: bidask v. 2.1.4 (Ardia et al., 2024), moments v. 0.14.1 (Komsta & Novomestky, 2022), pcaMethods v. 2.0.0 (Stacklies et al., 2007), PerformanceAnalytics v. 2.0.8 (Peterson & Carl, 2024), quantmod v. 0.4.28 (Ryan & Ulrich, 2025), slider v. 0.3.2 (Vaughan, 2024), tidyverse v. 2.0.0 (Wickham et al., 2019), and zoo v. 1.8.14 (Zeileis & Grothendieck, 2005).

Additionally, I used Python 3.15.3 (Python Software Foundation, 2025) and the following packages: numpy (Harris et al., 2020), pandas (The pandas development team, 2020), matplotlib (Hunter, 2007), and scipy (Virtanen et al., 2020).