dxFeed Faces of the Crypto Market Index Family Methodology

dxFeed Index Management Team <im@dxfeed.com>

July 21, 2023

Contents

1 Summary

dxFeed Faces of the Crypto Market™ (“FACE_n”) Index Family provides a way to uniquely characterize the cryptocurrency market by analyzing the underlying price dynamics. In particular, each index within the family corresponds to an independent risk source, and is represented as a portfolio of crypto assets. The index value expresses the effective growth rate of the resulting synthetic instruments.

FACE_1 characterizes “the market” more accurately than the simple mean or the market cap-weighted average of the crypto coins. It also explains a large (typically more than 60%) proportion of the total price variation. FACE_2 might be considered an “anti-market” portfolio, containing crypto assets with significantly different return structure. FACE_3–FACE_6 are additional risk sources, which have low inter-correlation with the members of the index family. FACE_1 and FACE_2, and, to some degree, FACE_3, appear to be persistent across time, whereas the composition of the remaining indices tends to be less stable. A close inspection of those, however, may reveal coins with significant idiosyncratic behavior and relatively high contributions to the overall market risk.

The suggested instruments may serve as especially useful benchmarks for portfolio managers, and standalone assets in diversification and hedging scenarios.

2 Index Model

2.1 Eigenfaces

A well-known computational experiment [1] applies the method of *Principal Component Analysis* (PCA) to a dataset of human faces by decomposing each face into a linear combination of prototypical *“eigenfaces”*. Figure 1 shows a random sample from the Olivetti dataset of faces[1] used in the experiment. It also shows the first 10 eigenfaces resulting from applying PCA.

(a) Random sample of 10 faces from the dataset

(b) First 10 eigenfaces

Figure 1: An illustration of applying PCA to the Olivetti dataset of faces

More technically, each i-th face can be represented as a p-dimensional vector $\mathbf{x}_i \in \mathbb{R}^p$, $i \in 1:n$, where p is the number of pixels in each image and x_{ij} is the intensity of the j-th pixel, $j \in 1:p$. Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ also be a matrix of all such faces. Each face is then a linear combination of pixel intensities: if $\mathbf{e}_j = (0,\dots,0,1,0,\dots 0)^\intercal$ is a vector such that the only non-zero element is 1 at the j-th position, then

$$\mathbf{x}_i = \sum_{j=1}^{p} x_{ij}\mathbf{e}_j, \quad i \in 1:n.$$

The set $\{\mathbf{e}_1,\dots,\mathbf{e}_p\}$ is also known as the “*canonical basis*”. PCA finds a new orthonormal basis $\{\mathbf{u}_1,\dots,\mathbf{u}_p\}$

[1]Courtesy of AT&T Laboratories Cambridge; see https://cam-orl.co.uk/facedatabase.html. Pre-processed by Sam Roweis; see https://cs.nyu.edu/~roweis/data.html

(*"principal component vectors"*), such that

$$\mathbf{x}_i = \sum_{j=1}^{p} z_{ij}\mathbf{u}_j, \quad i \in 1:n,$$

i.e., the i-th face is a linear combination of eigenfaces. Each eigenface (a p-dimensional vector, i.e., a picture) captures significant facial features, such that their linear combination with appropriate coefficients (*"scores"*) $z_{i1}, \dots, z_{ip}$ can yield all faces from the initial dataset $\mathbf{X}$. $\{\mathbf{u}_1, \dots, \mathbf{u}_p\}$ is chosen in the decreasing order of explained variance, so $\mathbf{u}_1$ captures the most pronounced features in all the faces, $\mathbf{u}_2$ adds some particular differences, and so on.

2.2 Principal Instruments

Similarly, PCA can be applied to market data, in particular to the set of returns for a group of instruments. Let C be a set of p chosen instruments (e.g., crypto coins). Suppose $\boldsymbol{R}_t = (R_{t1}, \dots, R_{tp})^\top$ is the vector of simple returns for the p instruments on day $t \in 1 : n$. Just as the i-th face can be expressed as a linear combination of p pixel intensities in the canonical basis, the t-th day may be characterized by returns for p instruments,

$$\boldsymbol{R}_t = \sum_{j=1}^{p} R_{tj}\mathbf{e}_j, \quad t \in 1:n.$$

PCA then finds a new orthonormal basis $\mathcal{U} = \{\mathbf{u}_1, \dots, \mathbf{u}_p\}$ such that each trading day can be expressed as a linear combination of *synthetic instruments' returns* $\mathcal{U}$ and some coefficients z_{tj} :

$$\boldsymbol{R}_t = \sum_{j=1}^{p} z_{tj}\mathbf{u}_j, \quad t \in 1:n.$$

The synthetic instruments are referred to as "eigen-instruments" or "principal instruments". Each principal instrument is an *independent risk source*. Principal components are "synthetic" in the sense that they are expressed in terms of the original instruments' returns, much like eigenfaces are expressed in terms of the original pixel space.

Remark. Since the same space is spanned by a vector and its negative, the sign for each $\mathbf{u}_j$ is chosen such that the sum of squared loadings is less than $1/2$ (NB: $\sum_{i=1}^{p} u_{ij}^2 = 1$), i.e.,

$$\sum_{\substack{i=1 \\ u_{ij}<0}}^{p} u_{ij}^2 < 0.5.$$

This maximizes the long positions' weight in the resulting portfolios.

2.3 Growth Rate of Principal Instrument

The index value expresses the effective growth rate of the chosen principal instrument, as a percentage. Fix k-th principal instrument $\mathbf{u}_k$. Its elements are the original instruments' scaled returns in specific proportion. $\mathbf{u}_k$ has L_2 unit norm. Re-normalizing it to L_1 norm,

$$\mathbf{w} = \frac{\mathbf{u}_k}{\sum_{j=1}^{p} |u_{kj}|}, \tag{1}$$

yields a vector of *portfolio weights*. Given the weights, the portfolio return on date t can be computed as

$$R_\Pi(t) = \boldsymbol{R}_t^\top \mathbf{w},$$

and the *effective growth rate* is calculated recursively as

$$
\begin{aligned}
G_\Pi^{(E)}(t) = R_\Pi^{(E)}(t) + 1 = \prod_{t'=1}^{t}(1 + R_\Pi(t')) = G_\Pi^{(E)}(t-1)(1 + R_\Pi(t)) \\
= G_\Pi^{(E)}(t-2)(1 + R_\Pi(t-1))(1 + R_\Pi(t)) = \dots .
\end{aligned}
$$

However, to compute the actual index value at time t, a closed-form approximation for log-returns is used:

$$
\begin{aligned}
\hat{G}_\Pi^{(E)}(t) &= \prod_{j=1}^{p} S_j^{w_j}(t) \\
\text{Index}(t) &= \frac{100}{\text{Divisor}} \hat{G}_\Pi^{(E)}(t),
\end{aligned}
$$

where $S_j(t)$ is the j-th component price at time t, Divisor is an index parameter chosen to achieve index value continuity between rebalancings, and initially equal to $\hat{G}_\Pi^{(E)}(t_0)$, where t_0 is the time when the index was first computed.

3 Component Selection

The set of index components C consists of top-p crypto coins by market cap at rebalancing date t_R. While forming the set, the following coins are excluded:

- Stablecoins, e.g., USDT.
- Coins that had no available data for their corresponding USD-like pair (BTC/USDT, ETH/USDT, etc.) for more than 1% of days.

  The remaining missing data was filled using linear interpolation.
- All "dead coins", i.e., severely depreciated coins with no significant trading activity after a certain date, e.g., LUNA.

4 Parameter Derivation

Weights PCA is applied to the parameter estimation dataset formed by computing daily simple returns for all p coins during a predefined 1-year training period. The weights for each index are then calculated using 1.

Symbols (Sources & Quote Currencies) For each component, the symbol corresponding to the pair maximizing the 1-year trading volume preceding t_R over all USD-like quote instruments and exchanges is selected.

Divisor At each rebalancing at time t_R, the new divisor is computed as

$$
\text{Divisor} = \frac{100}{\text{Index}'(t_R)} \hat{G}_\Pi^{(E)}(t_R),
$$

where $\text{Index}'(t_R)$ is computed using the old parameters, and $\hat{G}_\Pi^{(E)}(t_R)$ using the new ones.

5 Rationale and Interpretation

In the discussion below, sample data helps to illustrate the general principles and some of the important statistical properties of the constructed indices. In particular, a time interval from 2022-01-01 to 2022-12-

31 is used for parameter estimation ("train" dataset), and for testing the following 6 months, from 2023-01-01 to 2023-06-30, are utilized ("test" dataset). The properties and principles discussed are expected to persist between periodic index rebalancings, although the particular numeric values will, of course, vary.

5.1 Proportion of Variance Explained and the Market Principal Coin

PCA effectively decomposes the selected set of cryptocurrencies into independent risk sources (*principal coins*). Each principal coin "explains" the overall market dynamics to some degree, as measured by the *"Proportion of Variance Explained"* (PVE) metric: if $\hat{\mathrm{D}}\mathbf{z}_j$ is the sample variance of the j-th principal coin scores, then its PVE is

$$\mathrm{PVE}_j = \frac{\hat{\mathrm{D}}\mathbf{z}_j}{\sum_{\ell=1}^{p} \hat{\mathrm{D}}\mathbf{z}_\ell}.$$

- Typically, the very first principal coin would explain a large proportion of market variability, hence it is appropriate to interpret it as *"the market"*.
- The other components may or may not have a significant PVE value, and it is more difficult to attribute economic meaning to them. However, some insights may be gained by analyzing the elements of the principal coin vectors $\{\mathbf{u}_2, \ldots, \mathbf{u}_p\}$ (see below).

Figure 2 illustrates a possible distribution of PVE over the principal components. PC_1 can be seen to explain about 60% of the total variance, whereas PC_2 explains as little as 4%. It takes a linear combination of the first 20 components to explain 90% of the total variance accounting for 40% of the total number of coins.

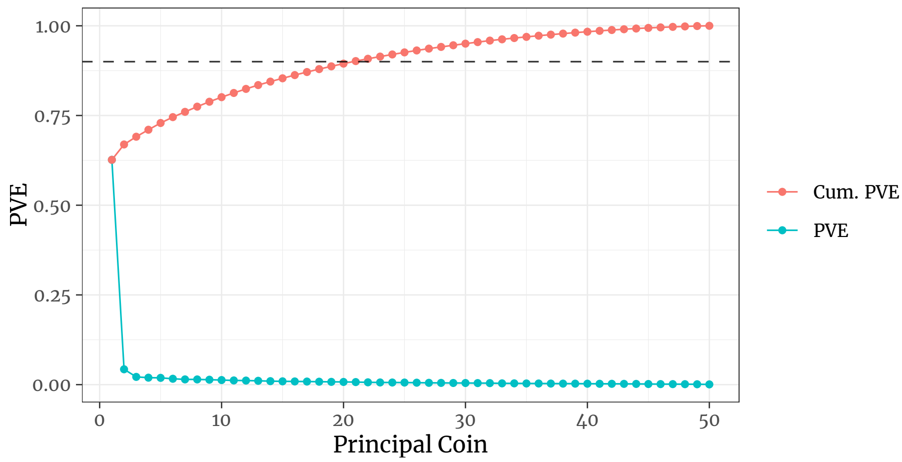


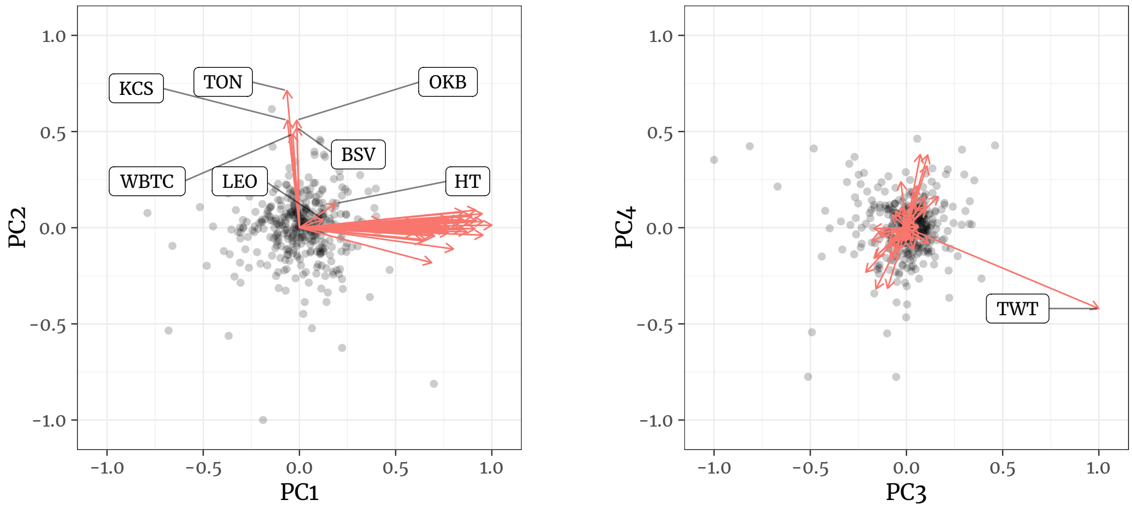
Figure 2: Illustration of PVE and cumulative PVE

5.2 The Biplots and the First Anti-Market Coin

One way to interpret the principal coins is to examine the plots of the original axes in the coordinate system defined by a pair of the principal coins (*"biplots"*), normalized to have unit standard deviation and confined within the $[-1, 1]$ interval. Original axes overlapping along a principal coin axis indicate that the corresponding coins can be explained by that single principal coin.

- Figure 3a is a boxplot of PC_1 vs. PC_2. It suggests the presence of a secondary set of coins, in addition to the set of coins forming "the crypto market", as represented by the bundle of axes along PC_1. This secondary set has a very different return structure than that of the rest of the coins (TON, KCS, ...); these would be the axes stretching along PC_2. That set of coins forms the second portfolio, an alternative to the "mainstream" market.

- Principal coins 3 and above are more difficult to interpret, a illustrated by Figure 3b. Depending on the time period, certain groups of coins may be more pronounced than the rest, but the effect isn't as persistent as in the case of the first to PCs.



(a) A biplot of PC_1 vs. PC_2 (b) A biplot of PC_3 vs. PC_4

Figure 3: Sample biplots

It is possible to attribute at least some financial meaning to PC_2: OKB (OK Exchange token), KCS (KuCoin Exchange token) and HT (Huobi Exchange token) are exchange utility tokens. WBTC (BTC wrapped for Ethereum) and BSV (BCH fork) are related to Bitcoin. However, TON (The Open Network coin) does not fit into any of these categories. In addition, not all exchange tokens are on the list (e.g., BNB, CRO). This reaffirms the fact that the principal coin portfolios cannot be composed based on common sense alone, and an analysis of instrument dynamics is necessary.

In this particular dataset, TWT seems to be "driving" PC_3 as discovered by the algorithm from the return structure during the parameter estimation period. This effect may or may not persist during later periods. However, correlation analysis shows that PC_3 has insignificant correlations with three other principal coins and a low correlation with the crypto market during the testing period (see below).

5.3 Principal Coin Loadings

The actual portfolio weights can be obtained by examining the loadings u_{ij}, as illustrated by Figure 4. The roughly equal loading values of PC_1 support the hypothesis that it corresponds to "the market" principal coin. However, some components have near-zero values (e.g., BSV, HT, ...), which suggests that they should not be a part of the market portfolio. Instead, their loadings are very pronounced in PC_2, which, again, supports the anti-market portfolio's hypothesis. The remaining principal coins are more difficult to interpret.

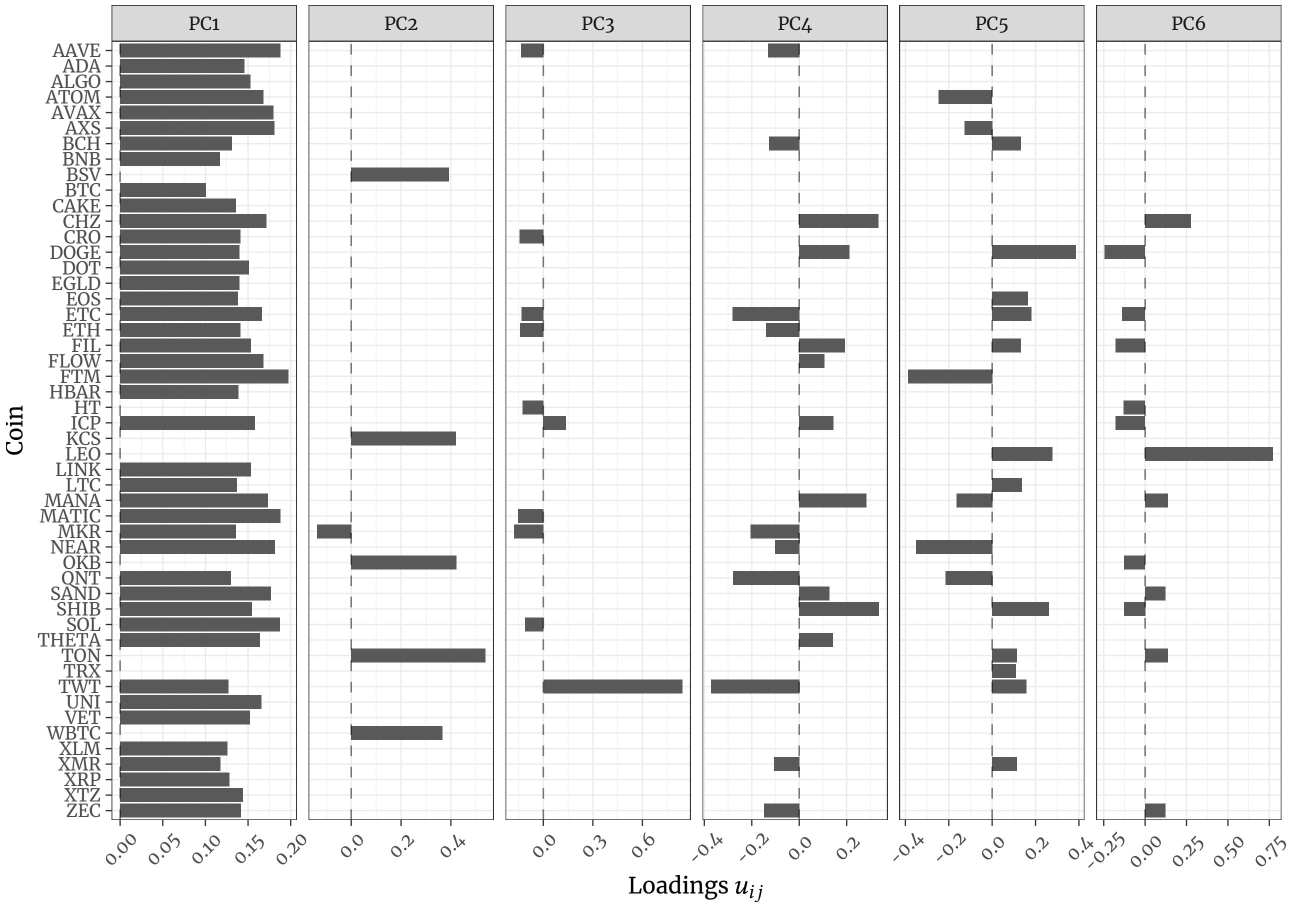


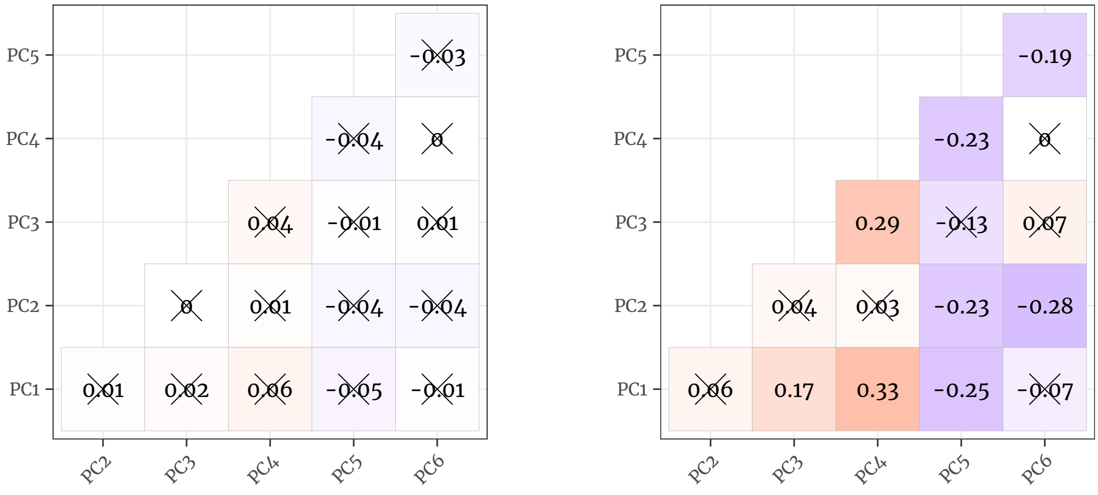
Figure 4: An illustration of component loadings u_{ij}. Values of $|u_{ij}| \leq 0.1$ are set to 0. Notice the different scales on the horizontal axis.

5.4 Return Correlation

One important property of PCA is that the obtained basis vectors $\{\mathbf{u}_1, \ldots, \mathbf{u}_p\}$ are orthogonal, implying that principal coin returns would have correlation numerically equal to zero. This property persists to some degree during the test period. As Figure 5a demonstrates, correlations are insignificant when computed over the whole period (train + test). However, they do become significant for some principal coins if computed over the test period only. Remarkably, PC_1 and PC_2 remain uncorrelated during the test period. Figure 6 gives a hint of how the property of uncorrelatedness deteriorates.

This finding may suggest that all principal coins above the first one are in a sense "anti-market", not just PC_2. Each principal coin provides a unique market characterization, although with diminishing explanatory power. In addition, insignificant inter-correlations or negative correlation with the market may be very desirable properties for portfolio construction.

How the principal coins are actually correlated depends on a particular index configuration and may differ from one rebalancing to another; the general idea, however, holds.



(a) A correlation matrix of principal coin returns over the whole period (train + test)

(b) A correlation matrix of principal coin returns over the test period only

Figure 5: Principal coins correlation matrices. Crossed out values indicate correlations insignificant at $p = 0.05$.

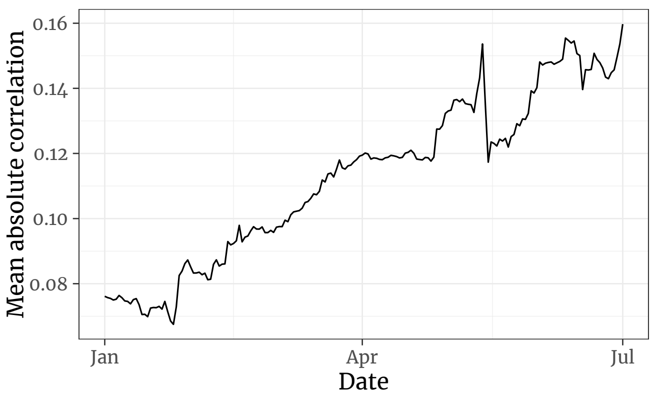


Figure 6: 6-months sliding mean absolute correlation of PC_1–PC_6 returns. Fisher z-transformation is applied before and after obtaining the mean. The initial correlation is not zero because the time frame is shorter than the PCA parameter estimation period (6 months vs. 1 year).

5.5 Composition Stability

Given a fixed set of cryptocurrencies, the obtained principal coin loadings (and hence the weights) are remarkably stable over time for PC_1, PC_2 and to some extent, PC_3. The remaining principal coins tend to be less stable and significantly noisier. See Figure 7 for details.

Figure 7: Rolling analysis of PCA loadings per principal coin over time. 1 year of daily simple returns is used to estimate the PCA loadings, starting from 2022-01-01. The procedure is repeated for the 6 following months (from 2023-01-01 to 2023-06-30), each day. Different colors correspond to different source coins.

That being said, the component set C does change significantly from rebalancing to rebalancing—this is because the set of top market cap coins varies over time.

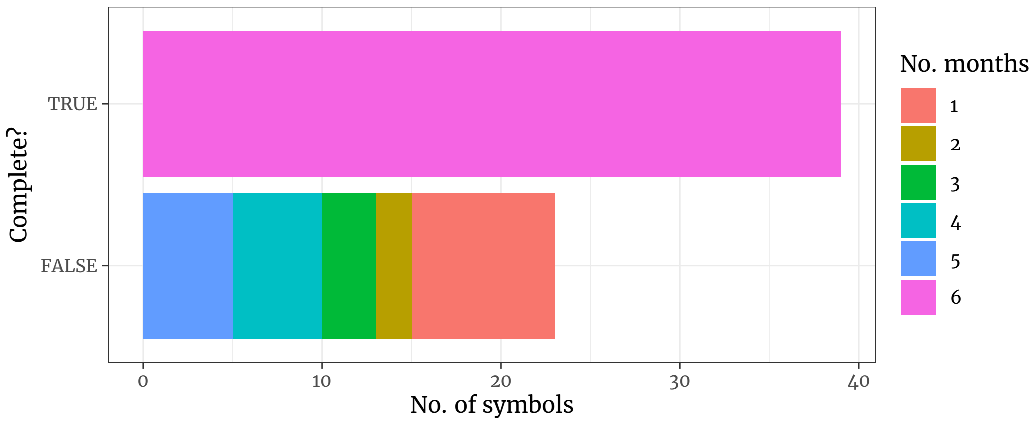


Figure 8: Illustration of PVE and cumulative PVE

5.6 Comparison with Standard Portfolios

As expected from the loading distributions (Figure 4), the PCA-obtained portfolios cannot be reduced to either equi-weighted or market cap-weighted portfolios, as Figure 9 demonstrates. The mcap-weighted portfolio is dominated by BTC and ETH. The equi-weighted portfolio resembles PC_1 except BSV, HT, and other coins, downweighted by the optimisation procedure. This equi-weighted portfolio is therefore a valuable addition to the standard set of widely used benchmark portfolios.

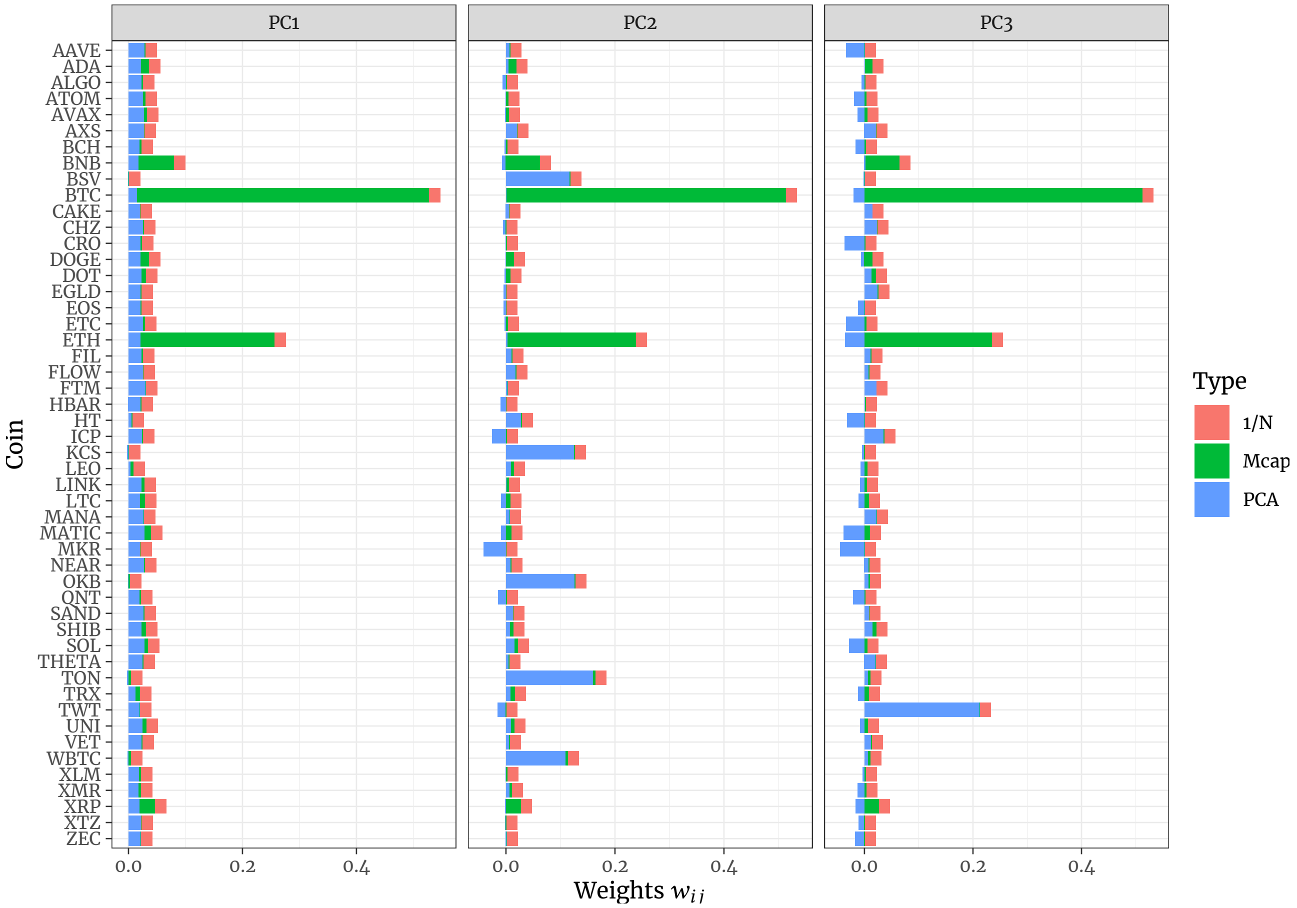


Figure 9: Component weights of the first three principal coins

5.7 Comparative Backtesting

It is instructive to compare both in- and out-of-sample performance of the principal coins vs. the standard equi-weighted and market cap-weighted portfolios.

- As Figure 10 illustrates, PC_1's performance is close to the market cap portfolio, as would be expected from the weight distribution, but the discrepancy between the two grows to more than 10% by the end of the period. This supports the idea that PC_1 can be viewed as yet another characterization of the market, and, according to the analysis, a better one than the simple equi-weighted portfolio.
- Other portfolios demonstrate unique dynamics, with some (e.g., PC_4) even having the growth rate significantly higher than 100% by the end of the period.

The observed price dynamics of the principal coins together with near-zero correlations suggest that the obtained portfolios may be used for diversification, hedging and similar scenarios.

Figure 10: Comparative backtesting of principal coins vs. standard portfolios. Dashed vertical lines indicate the beginning of the testing period.

6 Lifecycle & Maintenance

6.1 Rebalancing

The indices' composition is reviewed periodically—see the accompanying factsheet for details. The rebalancing might occur as a result of such a review.

Rebalancing procedures yield a new set of components C and their corresponding parameters (weights and symbols); see Sections 3 and 4.

6.2 Symbol Removal

If no rates can be fetched for component c^* due to the removal of the corresponding symbol, the default behavior is to use its last known value to compute the index ("last observation carried forward"). The index composition will be updated before or at the next scheduled rebalancing, at the discretion of the steering committee as follows:

1. Set $C' = C \setminus \{c^*\}$.
2. Renormalize the weights as $\text{Weight}'_{c'} = \text{Weight}_{c'} / \sum_{c \in C'} \text{Weight}_c$, for all $c' \in C'$.
3. Rebalance the index as in Section 6.1 using $\text{Weight}'_{C'}$, $\text{Symbol}_{C'}$.

References

[1] Matthew Turk, Alex Pentland; *Eigenfaces for Recognition.* J Cogn Neurosci 1991; 3 (1): 71–86. doi: https://doi.org/10.1162/jocn.1991.3.1.71

Disclaimer

Devexperts Inc. ("dxFeed") might receive compensation for licensing its indices to third parties and providing custom calculation services. dxFeed products are governed by the terms and conditions of the agreements under which they are provided. A license is required from dxFeed to display, create derivative works of, and/or distribute any product or service that uses, is based upon, and/or refers to any dxFeed index data.

It is not possible to invest directly in an index. Exposure to an asset class represented by an index is available through investable instruments based on that index. dxFeed makes no assurance that investment products based on the index will accurately track index performance or provide positive investment returns. dxFeed is not an investment advisor and makes no representation regarding the advisability of investing in any investment fund or other investment vehicle, including those based on dxFeed products or any documents and statements found in this document. Prices for dxFeed indices are calculated by dxFeed based on the prices of the index's individual constituents as set by their primary exchange or source. Prices are received by dxFeed either directly or from one of its third-party vendors. Vendors receive the prices from the primary exchanges.

Charts and graphs are provided for illustrative purposes only. The charts and graphs may reflect hypothetical historical performance. All information presented before the launch date is back-tested. Back-tested performance is hypothetical, not actual performance. The back-test calculations are based on the same methodology in effect after the indices' official launch. Back-tested performance reflects the application of an index's methodology and selection of index constituents. It adds the benefit of hindsight and knowledge of factors that may have positively affected its performance. Back-testing cannot account for all financial risks affecting results and may be considered to reflect survivor/look-ahead bias. Actual returns may differ significantly from, and be lower than, back-tested returns. Past performance is not an indication or guarantee of future results. Except for certain custom index calculation services, all information provided by dxFeed is impersonal. It is not tailored to the needs of any person, entity, or group of persons.

The materials contained in this document have been prepared solely for informational purposes based upon information generally available to the public, retrieved from sources believed to be reliable. No content within these materials (including index data, ratings, analysis, research, valuations, model, software, or other application or output therefrom), or any part thereof ("Content") may be modified, reverse-engineered, reproduced, or distributed in any form or by any means, nor copied or stored in a database or retrieval system, without dxFeed's prior written permission.

dxFeed does not assume any obligation to update the Content following publication in any form or format. The Content shall not be used for any unlawful or unauthorized purposes. dxFeed and its third-party data providers and licensors (collectively "dxFeed Parties") do not guarantee the Content's accuracy, completeness, timeliness, or availability. dxFeed is not responsible for any errors or omissions, regardless of the cause, for the results obtained from the Content's use.

The Content is provided on an "as is" basis. dxFeed disclaims all express or implied warranties, including, but not limited to, any warranties of merchantability or fitness for a particular purpose or use, freedom from bugs, software errors or defects, and warranties of the Content's uninterrupted functioning or operation with any software or hardware configuration.

In no event shall dxFeed be liable to any party for any direct, indirect, incidental, exemplary, compensatory, punitive, special, or consequential damages, costs, expenses, legal fees, or losses (including, but not limited to, lost income or lost profits and opportunity costs) in connection with any use of the Content, even if advised of the possibility of such damages.