# CRIX an Index for cryptocurrencies

Simon Trimborn [a,b,*], Wolfgang Karl Härdle [a,c]

[a] *Humboldt-Universität zu Berlin, C.A.S.E. - Center for Applied Statistics and Economics, Spandauer Str. 1, 10178 Berlin, Germany*
[b] *Department of Statistics and Applied Probability, National University of Singapore, 21 Lower Kent Ridge Rd, Singapore 119077, Singapore*
[c] *SKBI School of Business, Singapore Management University, 50 Stamford Road, Singapore 178899, Singapore*

## ARTICLE INFO

## ABSTRACT

The cryptocurrency market is unique on many levels: Very volatile, frequently changing market structure, emerging and vanishing of cryptocurrencies on a daily level. Following its development became a difficult task with the success of cryptocurrencies (CCs) other than Bitcoin. For fiat currency markets, the IMF offers the index SDR and, prior to the EUR, the ECU existed, which was an index representing the development of European currencies. Index providers decide on a fixed number of index constituents which will represent the market segment. It is a challenge to fix a number and develop rules for the constituents in view of the market changes. In the frequently changing CC market, this challenge is even more severe. A method relying on the AIC is proposed to quickly react to market changes and therefore enable us to create an index, referred to as CRIX, for the cryptocurrency market. CRIX is chosen by model selection such that it represents the market well to enable each interested party studying economic questions in this market and to invest into the market. The diversified nature of the CC market makes the inclusion of altcoins in the index product critical to improve tracking performance. We have shown that assigning optimal weights to altcoins helps to reduce the tracking errors of a CC portfolio, despite the fact that their market cap is much smaller relative to Bitcoin. The codes used here are available via www.quantlet.de.

## 1. Introduction

More and more companies have started offering digital payment systems. Smartphones have evolved into a digital wallet, telephone companies offer banking related services: clear signal that we are about to enter the era of digital finance. In fact we are already acting inside a digital economy. The market for e-$x$ ($x$ = "finance", "money", "book", you name it …) has not only picked up enormous momentum but has become standard for driving innovative activities in the global economy. A few clicks at $y$ and payment at $z$ brings our purchase to location $w$. Own currencies for the digital market were therefore just a matter of time. Due to organizational difficulties the idea of the Nobel Laureate Hayek, see Hayek (1990), of letting companies offer concurrent currencies seemed for a long time scarcely feasible, but the invention of the *Blockchain* has made it possible to bring his vision to life. Cryptocurrencies (CCs) have surfaced and opened up an angle towards this new level of economic interaction. Since the appearance of Bitcoins, several new CCs have spread through the Web and offered new ways of proliferation. Even states accept them as a legal payment method or part of economic interaction. E.g., the USA classifies CCs as commodities, Kawa (2015), and lately Japan announced that they accept them as a legal currency, EconoTimes (2016). Obviously, the crypto market is fanning out and shows clear signs of acceptance and deepening liquidity, so that a closer look at its general moves and dynamics is called for.

* Corresponding author at: Department of Statistics and Applied Probability, National University of Singapore, 21 Lower Kent Ridge Rd, Singapore 119077, Singapore.
*E-mail addresses:* simon.trimborn@nus.edu.sg (S. Trimborn), haerdle@hu-berlin.de (W.K. Härdle).

The transaction graph of Bitcoin (BTC), the Blockchain, has received much attention, see e.g. Ron and Shamir (2013) and Reid and Harrigan (2013). Even the economics of BTC has been studied, e.g. Bolt and Oordt (2016) and Kristoufek (2015). To our best knowledge, the development of the entire CC market has not been studied so far, only subsamples have been taken into account. Wang and Vergne (2017) studied the variations of 5 CCs. Elendner et al. (2017) analyzed the top 10 CCs by market capitalization and found that their returns are weakly correlated with each other. Furthermore, a Principal Component (PC) Analysis, carried out in the same reference, showed 7 out of 10 PC were necessary to describe more than 90% of the variance. These findings indicate the price evolution of CCs is very different from each other. This brings us to the conclusion that BTC, even though it dominates the market in terms of its market capitalization, cannot lead the direction of the market. The movements of other CCs are important too, when one analyzes the market. Having a closer look at the different CCs, it becomes obvious they have different kind of missions and technical aspects. Bitcoin pioneered as the token of the first decentralized, distributed ledger, giving start to multiple interpretations of its nature and purpose: new type of currency, commodity (like gold), alternative asset or innovative technology. The currently second most important CC by market capitalization—Ethereum – was created with a particular goal in mind – to power the blockchain based Ethereum platform for company building (DAO) and smart contract implementation. This idea triggered an unprecedented interest as it allowed companies to enter the field without creating their own blockchain ecosystem. Newcomers could benefit from the existing supporters of the respective platform, which allowed faster entry, adoption and operation. Other CCs, like Ripple (XRP), are intended to fuel the transaction network bridging traditional markets (banks) and the crypto ecosystem. Ripple also became one of the first successful cases of pre-emitted CC, abandoning the idea of decentralization. Since the appearance of BTC many technological advancements took place. Some CCs are designed for faster (or even immediate) transactions, like Litecoin (LTC), some are more efficient energy-wise, like DASH. Many embraced different hashing algorithms, altering the mining process, like Monero. Long ASIC domination is being disrupted, Proof-of-work is replaced by Proof-of-Stake, new ways to motivate those providing computational power are introduced. Regardless the type of CC, one witnesses a new kind of transaction network with a different approach for fees and handling of trust issues. The intended and actual usage can be interpreted as the business model of the different CCs and the participation in either CC can give advantages over others, White (2014).

In the first month of 2017, CCs other than BTC (altcoins) showed a strong gain in their market capitalization, reducing the dominance of BTC in the market. The finding of very different movements of CCs and the stronger position of alternative CCs in the market infers the necessity of a market index for the CC market for tracking the market movements. Comparing CCs against a market index answers economic questions like which business model is more successful than another one, gained recently compared to other CCs, drives the success of the market, is more established. Comparing a CC market index against other market indices answers economic and financial questions like which market proxy is more volatile, has more tail risk, attracts more investments. We construct CRIX, a market index (benchmark) which will enable each interested party to study the outlined economic questions, the performance of the CC market as a whole or of single CCs. Studying the stochastic dynamics of CRIX will allow a la limite to create ETFs or contingent claims.

Many index providers construct their indices with a fixed number of constituents, see e.g. FTSE (2016), S&P (2014) and Deutsche Boerse AG (2013). If the respective index is intended to be a proxy for the performance of a market, this requires huge trust from economists and investors into the choice of the index constituents by the index provider. On the other hand, the CRSP index family, derived for the US market, CRSP (2015), has no boundary on the number of index constituents. The number of constituents is reviewed daily and adjusted until the index members cover a predefined share of the market capitalization. Such a dynamic methodology is important in the market of CCs since the number of CCs changes daily. Additionally the market value of CCs often changes frequently, which increases the market volatility and therefore the need for considering such a CC for the representation of the market. Our intention is extending the idea behind the CRSP indices. Our first goal is constructing a methodology for CRIX which relies on model selection criteria to receive a proxy for the market and to replace the trust problematic with a statistical methodology. The resulting methodology is dynamic in the number of index constituents, like the CRSP indices. By this method only CCs which add informative value to the index are considered, which makes it representative. If more CCs than BTC are necessary to fulfill this requirement, they will be added. However we are concerned with the dominance of BTC in an index solely relying on market capitalization. Thus we introduce a second weighting scheme based on weighting by trading volume. Due to the usage of trading volume, the respective index is constructed in terms of trading focus. If the market participants focus more on altcoins than on BTC, these receive a higher weight. On the other hand, if the market focus is truly on BTC, it will receive a high weight in either index. Our second goal, constructing an investable index will be fulfilled by the methodology itself due to having a sparse index, only consisting of actively traded CCs in a market with low transaction costs. Note that due to the low transaction costs in the CC market, a dynamic methodology creates low additional costs. Additionally to the methodology ensuring an investable index, the proposed trading volume weighting scheme further supports this goal.

Investing into an ETF composed of the constituents of CRIX implies some differences compared to traditional index investing. In the traditional setting only the constituents are reviewed and replaced on the review date – if necessary – according to the index rules. In dynamic index investing the constituents are also reviewed for their number. This requires the manager of the fund to buy and sell more assets on the review date. In a market with high transaction costs, this approach is more costly. But the market of CCs has very low transaction costs, thus this problem will not occur in this market.

To compute CRIX, the differences in the log returns of the market against a selection of possible indices is evaluated. The results show, that the AIC works well to evaluate the differences. It penalizes the index for the number of constituents. For the calculation of the respective likelihoods, a non-parametric approach using the Epanechnikov (1969) kernel is applied. The proof for the impact of the value of an asset in the market on the AIC method is given, thus a top-down approach is applied to select the assets for the benchmarks to choose from, where the sorting depends on either market cap or trading volume. The number of constituents is

recalculated quarterly to ensure an up-to-date fit to the current market situation. With CRIX one may study the contingent claims and the stochastic nature of this index, Chen et al. (2017), or study the CC market characteristics against traditional markets, Härdle and Trimborn (2015).

This paper is structured as follows. Section 2 introduces the topic and reviews the basics of index construction. In Section 3 the method for dynamic index construction for CRIX is described and Section 4 introduces the remaining rules for CRIX. Section 5 describes further variants to create a CRIX family. Their performance is tested in Section 6. In Sections 7 and 8 the new method is applied to the German and Mexican stock markets to check the performance of the methodology against existing indices. The codes used to obtain the results in this paper are available via www.quantlet.de.

## 2. Index construction

The basic idea of any price index is to weight the prices of its constituent goods by the quantities of the goods purchased or consumed. The Laspeyres index takes the value of a basket of $k$ assets and compares it against a base period:

$$P_{0t}^{L}(k) = \frac{\sum_{i=1}^{k} P_{it}Q_{i0}}{\sum_{i=1}^{k} P_{i0}Q_{i0}} \tag{1}$$

with $P_{it}$ the price of asset $i$ at time $t$ and $Q_{i0}$ the quantity of asset $i$ at time 0 (the base period). For market indices, such as CRSP, S&P500 or DAX, the quantity $Q_{i0}$ is the number of shares of the asset $i$ in the base period. Multiplied with its corresponding price, the market capitalization results, hence the constituents of the index are weighted by their market capitalizations. These indices are often referred to as benchmarks for their respective market. We define the term benchmark:

**Definition 1.** A benchmark is a measure which consists of a selection of CCs that are representing the market.

But markets change. A company which was representative for market developments yesterday might no longer be important today. On top of that, companies can go bankrupt, a corporation can raise the number of its outstanding shares, or trading in it can become infrequent. All these situations must produce a change in the index structure, so that the market is still adequately represented. Hence companies have to drop out of the index and have to be replaced by others. The index rules determine in which cases such an event happens. The formula of Laspeyres (1) cannot handle such events entirely because a change of constituents will result in a change in the index value that is not due to price changes. Therefore, established price indices like DAX or S&P500, see Deutsche Boerse AG (2013) and S&P (2014) respectively, and the newly founded index CRIX($k$), a CRyptocurrency IndeX, thecrix.de, use the adjusted, formula of Laspeyres,

$$\text{CRIX}_t(k, \beta) = \frac{\sum_{i=1}^{k} \beta_{i,t_l^-} P_{it} Q_{i,t_l^-}}{Divisor(k)_{t_l^-}} \tag{2}$$

with $P$, $Q$ and $i$ defined as before, $\beta_{i,t_l^-}$ the adjustment factor of asset $i$ found at time point $t_l^-$, $l$ indicates that this is the $l$th adjustment factor, and $t_l^-$ the last time point when $Q_{i,t_l^-}$, $Divisor(k)_{i,t_l^-}$ and $\beta_{i,t_l^-}$ were updated. In the classical setting, $\beta_{i,t_l^-}$ is defined to be $\beta_{i,t_l^-} = 1$ for all $i$ and $l$. Anyhow, some indices use $\beta_{i,t_l^-}$ to achieve maximum weighting rules, e.g. Deutsche Boerse AG (2013) and MEXBOL (2013). The *Divisor* ensures that the index value of CRIX has a predefined value on the starting date. It is defined as

$$Divisor(k, \beta)_0 = \frac{\sum_{i=1}^{k} \beta_{i0} P_{i0} Q_{i0}}{\text{starting value}}. \tag{3}$$

The starting value could be any possible number, commonly 100, 1000 or 10000. It ensures that a positive or negative development from the base period will be revealed. Whenever changes to the structure of CRIX occur, the *Divisor* is adjusted in such a way that only price changes are reflected by the index. Defining $k_1$ and $k_2$ as number of constituents, it results

$$\frac{\sum_{i=1}^{k_1} \beta_{i,t_{l-1}^-} P_{i,t-1} Q_{i,t_{l-1}^-}}{Divisor(k_1, \beta)_{t_{l-1}^-}} = \text{CRIX}_{t-1}(k_1, \beta) = \text{CRIX}_t(k_2, \beta) = \frac{\sum_{j=1}^{k_2} \beta_{j,t_l^-} P_{j,t} Q_{j,t_l^-}}{Divisor(k_2, \beta)_{t_l^-}}. \tag{4}$$

In indices like FTSE, S&P500 or DAX the number of index members is fixed, $k_1 = k_2$, see FTSE (2016), S&P (2014) and Deutsche Boerse AG (2013). As long as the goal behind these indices is the reflection of the price development of the selected assets, this is a straightforward approach. But, e.g., DAX is also meant to be an indicator for the development of the market as a whole, see Janßen and Rudolph (1992). This raises automatically the question of whether the included assets and the weighting scheme are representing the market. Since the constituents are chosen using a top-down approach, meaning that the biggest companies by market capitalization are included, the intuitive answer is yes. But it leaves a sour taste that additional assets may describe the market more appropriately. Furthermore different weighting schemes provide another view on the market. One may object by referring to total market indices like the Wilshire 5000, S&P Total Market Index or CRSP U.S. Total Market Index, see Wilshire Associates (2015), S&P (2015) and CRSP (2015), that are providing a full description. But financial practice has shown that smaller indices like DAX30 and S&P500 receive more attention in evaluating the movements of their corresponding markets, probably because they are easier to invest in due to the smaller number of constituents. It is therefore appealing to know which are the representative assets in a market and which smaller number of index constituents eases the handling of a tracking portfolio. Additionally, one may be concerned that an index would include illiquid and non-investable assets which makes the management of a tracking portfolio even more difficult. Fig. 1 shows that this is indeed a problem in the CC market. Some CCs have a fairly high market capitalization while their respective trading volume

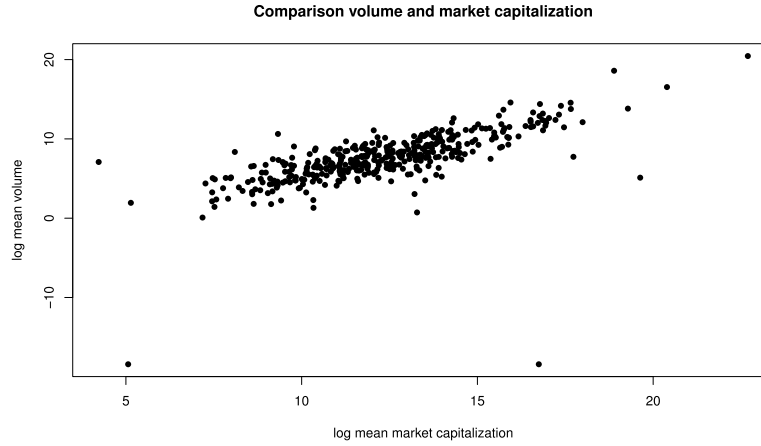**Comparison volume and market capitalization**



**Fig. 1.** Comparison of the log mean trading volume and log mean market capitalization, both measured in USD, for all CCs in the dataset over the time period 20140401–20170325. VolMarketCapComparison.

**Table 1**
Weighting schemes for derivation of CRIX.

|  | Market cap weighting | Liquidity weighting |
|---|---|---|
| $\beta_{i,t_l^-}$ | 1 | $\frac{Vol_{i,t_l^-}}{P_{i,t_l^-} Q_{i,t_l^-}}$ |

is very low. This is problematic, because an asset which is not frequently traded cannot add enough information to a market index to display market changes and is difficult to trade for an investor. Hence, one goal behind constructing CRIX is making it investable by concentrating on liquid CCs:

**Definition 2.** Between investment portfolios with equal performance, the one with the least assets is preferable.

We react to the goals and problems in two ways: First, these thoughts raise the question which value of $k$ is "optimal" for building an investable benchmark for the market. Additionally, especially young and innovative markets may change their structure over time. Therefore, a quantification of an accurate CC benchmark with sparse number of constituents is asked for. Since the CC market shows a frequently changing market structure with a huge number of illiquid CCs, a time varying index selection structure is applied. The later described selection method omits illiquid CCs by construction, because only CCs who show changes in their return series can be selected to be added to CRIX by the method. Due to the low transaction costs in this market, a dynamic methodology is applicable since it does not raise the costs of restructuring a tracking portfolio too much. Secondly, we apply two kind of weighting schemes, Table 1. We apply the classical setting to build a proper market index which is only flexible in terms of the dynamic constituents and tackles the illiquidity issue due to the applied selection method. The liquidity weighting allows one to weight CCs higher, which are more traded relative to their market capitalization and therefore implicitly acquire more financial attention. This weighting scheme bails (2) down to weighting the price development by their trading volume,

$$\text{LCRIX}_t(k, \beta) = \frac{\sum_{i=1}^{k} \frac{Vol_{i,t_l^-}}{P_{i,t_l^-} Q_{i,t_l^-}} P_{it} Q_{i,t_l^-}}{Divisor(k)_{t_l^-}} = \frac{\sum_{i=1}^{k} \frac{Vol_{i,t_l^-}}{P_{i,t_l^-}} P_{it}}{Divisor(k)_{t_l^-}}.$$ (5)

The latter is referred to as Liquidity CRIX (LCRIX). This approach has the potential to diminish the influence of e.g. Bitcoin stronger than the market cap weighting, if the relation of trading volume to market cap is higher for other CCs. In Section 6 we show that LCRIX has a better mean directional accuracy than CRIX and puts more weight on altcoins, Table 8, therefore tackling the issue of BTC dominance when the actual trading amount suggests a different result.

## 3. Dynamic index construction

This section is dedicated to describing the composition rule which is used to find the number of index members—the spine of CRIX and LCRIX. Since CRIX will be a benchmark for the CC market, the dimension and evaluation of the market has to be defined:

**Definition 3.** The total market (TM) consists of all CCs in the CC universe. Its value is the combined market value of the CCs.

To compare the TM with a benchmark candidate, it will be normalized by a Divisor,

$$\text{TM}(K)_t = \frac{\sum_{i=1}^{K} P_{it} Q_{i,t_l^-}}{Divisor(K)_{t_l^-}}$$ (6)

with $K$ the number of all CCs in the CC universe. Note that no adjustment factor is used for $TM(K)_t$. For the volume weighting, the TM is defined as LTM respectively,

$$\text{LTM}(K)_t = \frac{\sum_{i=1}^{K} \frac{Vol_{i,t_l^-}}{P_{i,t_l^-} Q_{i,t_l^-}} P_{it} Q_{i,t_l^-}}{Divisor(K)_{t_l^-}}. \tag{7}$$

In the further explanations, the focus lies on the TM. However when LCRIX is derived, it is optimized against LTM. The results can be easily extended to the case of LTM. Further define the log returns:

$$\varepsilon(K)_t^{TM} = \log\{TM(K)_t\} - \log\{TM(K)_{t-1}\} \tag{8}$$

$$\varepsilon(k, \beta)_t^{CRIX} = \log\{CRIX(k, \beta)_t\} - \log\{CRIX(k, \beta)_{t-1}\}, \tag{9}$$

where $CRIX(k, \beta)_t$ is the CRIX with $k$ constituents at time point $t$.

The goal is to optimize $k$ and $\beta$ so that a sparse but accurate approximation in terms of

$$\min_{k,\beta} \|\varepsilon(k, \beta)\|^2 = \min_{k,\beta} \|\varepsilon(K)^{TM} - \varepsilon(k, \beta)^{CRIX}\|^2, \tag{10}$$

is achieved, where $\varepsilon(k, \beta)$ is the difference in the log returns of $TM(K)$ and $CRIX(k, \beta)$. A squared loss function is chosen in (10), since it heavily penalizes deviations.

Since the value of $TM(K)_t$ is unknown and not measurable due to a lack of information, the total market index will be defined and used as a proxy for the $TM(K)$. The definition is inspired by total market indices like CRSP (2015), S&P (2015) and Wilshire Associates (2015). They use all stocks for which prices are available.

**Definition 4.** The total market index (TMI) contains all CCs in the CC universe for which prices are available. The CCs are weighted by their market capitalization.

This changes (6) to

$$\text{TMI}_t(k_{max}) = \frac{\sum_{i=1}^{k_{max}} P_{it} Q_{i,t_l^-}}{Divisor(k_{max})_{t_l^-}}$$

with $k_{max}$ the maximum number of CCs with available prices and (10) to

$$\min_{k,\beta} \|\hat{\varepsilon}(k, \beta)\|^2 = \min_{k,\beta} \|\varepsilon(k_{max})^{TMI} - \varepsilon(k, \beta)^{CRIX}\|^2 \tag{11}$$

$$\text{s.t.: } 1 \leq k \leq k^u$$

$$k = k_1 + s \tag{12}$$

$$k^u \in [1, k_{max}]$$

$$s \in [1, k_{max} - k_1]$$

$$\beta^{1 \times k} = (1, \dots, 1, \beta_{k_1+1}, \dots, \beta_{k_1+s})^\top$$

$$\beta_{k_1+1}, \dots, \beta_{k_1+s} \in (-\infty, \infty),$$

where $\varepsilon(k_{max})^{TMI}$ are the log returns for TMI. In the derivation of LCRIX, the optimization is performed against LTMI and $\beta^{1 \times k} = (\beta_1, \dots, \beta_k, \beta_{k_1+1}, \dots, \beta_{k_1+s})^\top$ where $\beta_i = \frac{Vol_{i,t_l^-}}{P_{i,t_l^-} Q_{i,t_l^-}}$ for $i = 1, \dots, k_1$ and $\beta_{k_1+1}, \dots, \beta_{k_1+s} \in (-\infty, \infty)$.

Several constraints were introduced with (11). The parameters $\beta_{k_1+1}, \dots, \beta_{k_1+s}$ are included to evaluate if adding $s$ more assets to the index explains the difference between $\varepsilon(k_{max})^{TMI}$ and $\varepsilon(k, \beta)^{CRIX}$ better. The first $k$ assets $(k_1)$ will not be adjusted by a parameter, so no parameter estimation is necessary. This makes the first term a constant. The choice of $k_1$ is important since it defines the number of base CCs to be included in the index. The parameters of the next $s$ assets have to be estimated, so (2) becomes

$$\text{CRIX}_t(k, \beta) = \frac{\sum_{i=1}^{k_1} P_{it} Q_{i,t_l^-} + \sum_{j=k_1+1}^{k_1+s} \beta_{j,t_l^-} P_{jt} Q_{j,t_l^-}}{Divisor(k_1)_{t_l^-}}.$$

A number of criteria are applicable. Model selection (SC) criteria can be categorized by their property to be either asymptotic optimal or consistent in choosing the true model. In this context will be investigated: Generalized Cross Validation (GC), Generalized Full Cross Validation (GFC), Mallows' $C_p$, Shibata (SH), Final Prediction Error (FPE) and Akaike Information Criterion (AIC), all asymptotic optimal criteria under the assumption of Gaussian distributed residuals. Since CRIX is supposed to be a benchmark model, all possible models under certain restrictions for the number of parameters are included in the test set,

$$\Theta_{SC} = \{CRIX(k_1, \beta), CRIX(k_2, \beta), \dots\}, \tag{13}$$

where $k_1, k_2, \dots$ are predefined values and $SC \in \{GC, GFC, C_p, SH, FPE, AIC\}$. Recall that the intention behind CRIX is to discover under a squared loss function the best model to describe the data (benchmark), which supports the choice of an asymptotic optimal criteria. The GC criterion, see Craven and Wahba (1978), is defined as

$$\text{GC}\{\hat{\varepsilon}(k, \beta), s\} = \frac{T^{-1} \sum_{t=1}^{T} \hat{\varepsilon}(k, \beta)_t^2}{(1 - T^{-1}s)^2} \tag{14}$$

by assuming that $s < T$. One shall note that $s$ and not $k + s$ defines the number of variables to penalize for, since $k$ parameters are set to be 1 and need not be estimated. According to Arlot and Celisse (2010), the asymptotic optimality of GC was shown in several frameworks. The GFC, see Droge (1996):

$$\text{GFC}\{\widehat{\varepsilon}(k, \beta), s\} = T^{-1} \sum_{t=1}^{T} \widehat{\varepsilon}(k, \beta)_t^2 (1 + T^{-1}s)^2 \qquad (15)$$

is an alteration.

A further score, SH,

$$\text{SH}\{\widehat{\varepsilon}(k, \beta), s\} = \frac{T + 2s}{T^2} \sum_{t=1}^{T} \widehat{\varepsilon}(k, \beta)_t^2, \qquad (16)$$

was shown to be asymptotically optimal, Shibata (1981), and asymptotically equivalent to Mallows' $C_p$ and AIC.

Mallows (1973)' $C_p$:

$$C_p\{\widehat{\varepsilon}(k, \beta), s\} = \frac{\sum_{t=1}^{T} \widehat{\varepsilon}(k, \beta)_t^2}{\widehat{\sigma}(k, \beta)^2} - T + 2 \cdot s \qquad (17)$$

with $\widehat{\sigma}(k, \beta)^2$ the variance of $\widehat{\varepsilon}(k, \beta)$. $C_p\{\widehat{\varepsilon}(k, \beta), s\}$ tends to choose models which overfit and is not consistent in selecting the true model, see Mallick and Yi (2013), Woodroofe (1982) and Nishii (1984).

The FPE uses the formula

$$\text{FPE}\{\widehat{\varepsilon}(k, \beta), s\} = \frac{T + s}{(T - s)T} \sum_{t=1}^{T} \widehat{\varepsilon}(k, \beta)_t^2, \qquad (18)$$

see Akaike (1970)

So far, the discussed criteria depend on little data information. Just the squared residuals and, in the case of Mallows' $C_p$, the variance are taken into account. The AIC uses more information by depending on the maximum likelihood, derived by

$$L\{\widehat{\varepsilon}(k, \beta)\} = \max_{\beta} \prod_{t} f\{\widehat{\varepsilon}(k, \beta)_t\}, \qquad (19)$$

where $f$, in (21), represents the density of the $\widehat{\varepsilon}(k, \beta)_t$ over all $t$. The AIC is defined to be

$$\text{AIC}\{\widehat{\varepsilon}(k, \beta), s\} = -2 \log L\{\widehat{\varepsilon}(k, \beta)\} + s \cdot 2, \qquad (20)$$

Akaike (1998). If the true model is of finite dimension, then the AIC is not consistent, compare Hurvich and Tsai (1989). Shibata (1983) showed the asymptotic efficiency of Mallows' $C_p$ and AIC under the assumption of an infinite number of regression variables or an increasing number of regression variables with the sample size. Due to the usage of the density in deriving the AIC, it uses more information about the dataset. Considering that (10) implies the criteria are derived under an expected squared loss function,

$$\text{E}(\|\varepsilon(k, \beta)\|^2) = \int_{-\infty}^{\infty} \|\varepsilon(k, \beta)\|_2^2 f\{\varepsilon(k, \beta)\} d\varepsilon(k, \beta), \qquad (21)$$

the density, $f$, can be estimated different from the Gaussian distribution. Here, $f$ is estimated nonparametrically with an Epanechnikov kernel, since according to Härdle et al. (2004) the Epanechnikov (1969) kernel shows a good balance between variance optimization and numerical performance. In nonparametric estimation with an Epanechnikov kernel, Epa, the estimator of $f$ is derived by

$$\widehat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} \text{Epa}\left(\frac{x - x_i}{h}\right), \quad \text{Epa}(u) = \frac{3}{4\sqrt{5}}\left(1 - \frac{u^2}{5}\right)\mathbf{I}(|u| \leq \sqrt{5})$$

where $h$ is the bandwidth.

The bandwidth selection is performed with the plug-in selector by Sheather and Jones (1991) and further described in Wand and Jones (1994). The plug-in selector is derived under the loss function Mean Integrated Squared Error, MISE. Hall (1987) found that the Kullback–Leibler (KL) loss function for selecting the smoothing parameter of the kernel density is highly influenced by the tails of the distribution. Devroye and Györfi (1985) mention that Mean Integrated Error (MIE) is stronger affected than MISE by the tails of the distribution and Kanazawa (1993) claims that MIE shall be used if interest is in modeling the tails. Kanazawa (1993) investigates that the use of a Kullback–Leibler loss function would put more weight on the tails compared to MISE. Since this is not in our interest, the choice of the density smoothing parameter, $h$, is performed under MISE.

Due to the richer information basis of the AIC, we decide to use it as the selection criteria for CRIX. The choice is supported by an empirical analysis in Section 6.

To decide with the AIC which number $k$ should be used, a procedure was created which compares the squared difference between log returns of the TMI, see Definition 4, and several candidate indices,

$$\|\widehat{\varepsilon}(k_j, \beta)\|^2 = \|\varepsilon(k_{max})^{TMI} - \varepsilon(k_j, \beta)^{CRIX}\|^2, \qquad (22)$$

where $\varepsilon(k_j, \beta)^{CRIX}$ is the log return of CRIX version with $k_j$ constituents and $\widehat{\varepsilon}(k_j, \beta)$ is the respective difference. The candidate indices, $\text{CRIX}(k_j, \beta)$, have different numbers of constituents which fulfill $k_1 < k_2 < k_3 < \cdots$, where $k_j = k_1 + s(j - 1)$. Therefore, the

number of constituents between the indices are equally spaced. The procedure implies that the selection method evaluates if $s$ more assets add information to CRIX. If so, these assets are added to the intercept and the next $s$ assets are tested for. Assets with a higher market capitalization are expected to have a higher influence on the AIC, so the following theorem is formulated:

**Theorem 1.** *The rate of improvement of the AIC depends on the relative value of an asset in the market.*

The proof for Theorem 1 is given in the Appendix A.1, under the assumption of normally distributed error terms. Therefore, we will follow the common practise to include the assets with the highest market capitalization in the index,

$$\arg\max_i \sum_{j=1}^{k} P_{j,i,t_l^-} Q_{j,i,t_l^-}, \quad i \in \{1, \dots, K\}. \tag{23}$$

Thus, a top-down approach to decide about the number of index constituents is applied.

For the sorting of the index constituents by highest market capitalization, just the closing data of the last day of a month are used. We chose to do so, since the next periods CRIX will just depend on $Q_{i,t_l^-}$, (2), and not on data which lie further in the past. This is in line with the methodology of e.g. the DAX. For LCRIX, the CCs with the highest trading volume are chosen respectively,

$$\arg\max_i \sum_{j=1}^{k} Vol_{j,i,t_l^-}, \quad i \in \{1, \dots, K\}. \tag{24}$$

Since the differences between the TMI($k_{max}$) and CRIX($k_j, \beta$) are caused over time by the missing time series in CRIX($k_j, \beta$), the independence assumption of the $\hat{\varepsilon}(k_j, \beta)$ for all $j$ cannot be fulfilled by construction. But Györfi et al. (1989) give arguments that under certain conditions in case of nonparametric density estimation, the rate of convergence is essentially the same as for an independent sample. Summarizing the described procedure, results to:

1. At time point $T + 1$, construct TMI($k_{max}$)
2. Set $j = 2$
3. Construct CRIX($k_1, 1$) and CRIX($k_j, \beta$), $k_1 < k_2 < k_3 < \cdots$
4. Compute $\hat{\varepsilon}(k_j, \beta)$ and $\hat{\varepsilon}(k_1, 1)$
5. Kernel density estimation (KDE) for density $f(\hat{\varepsilon}(k_1, 1))$

   (a) Compute the log likelihood (20) for $\hat{\varepsilon}(k_j, \beta)$ with KDE for $\hat{\varepsilon}(k_1, 1)$.
   (b) Sum the log likelihoods

6. Derive AIC$\{\hat{\varepsilon}(k_j, \beta), k_j - k_1\}$ and AIC$\{\hat{\varepsilon}(k_1, 1), 0\}$
7. If $j = (k_{max} - k_1)/k_1$: stop, else jump to 3. and $j = j + 1$

The next section describes the further index rules for CRIX.

## 4. CRIX family rules

The constituents of the indices are regularly checked so that the corresponding index always represents its asset universe well. It is common to do this on a quarterly basis. In case of CRIX this reallocation is much faster. In the past, coins have shown a very volatile behavior, not just in the manner of price volatility. In some weeks, many occur out of nothing in the market and many others vanish from the market even when they were before very important, e.g., Auroracoin. This calls for a faster reallocation of the market benchmark than on a quarterly basis. A monthly reallocation is chosen to make sure that CRIX catches the momentum of the CC market well. Therefore, on the last day of every month, the CCs which had the highest market capitalization on the last day in the last month will be checked and the first $k$ will be included in CRIX for the coming month. Accordingly for LCRIX the ones with the highest trading volume are chosen.

Since a review of an index is commonly performed on a quarterly basis the number of index members of CRIX will be checked on a quarterly basis too. The described procedure from Section 3 will be applied to the observations from the last three months on the last day of the third month after the markets closed. The number of index constituents, $k$, will be used for the next three months. Thus, CRIX corresponds to a monthly rebalanced portfolio which number of constituents is reviewed quarterly.

It may happen that some data are missing for some of the analyzed time series. If an isolated missing value occurs alone in the dataset, meaning that the values before and after it are not missing, then Missing At Random (MAR) is assumed. This assumption means that just observed information cause the missingness, Horton and Kleinman (2007). The Last-Observation-Carried-Forward (LOCF) method is then applied to fill the gap for the application of the AIC. We did not choose a different approach since a regression or imputation method may alter the data in the wrong direction. By LOCF, no change is implied and the CC is not excluded. If two or more data are missing in a row, then the MAR assumption may be violated, therefore no method is applied. The corresponding time series is then excluded from the computation in the derivation period. If data are missing during the computation of the index values, the LOCF method is applied too. This is done to make the index insensitive to this CC at this time point. CRIX should mimic market changes, therefore an imputation or regression method for the missing data would distort the view on the market.

Before continuing, the described rules are summarized:

- Quarterly altering of the number of index constituents

- Monthly altering of the index constituents
- Model selection for index derivation with AIC
- Nonparametric estimation of the density
- Application of a top-down approach to select the assets for the subset analysis
- Application of LOCF if trading of an asset stops before next reallocation.

## 5. The CRIX family

Using the described methods and rules from above, three indices will be proposed. This indices provide a different look at the market.

1. CRIX/LCRIX: The first and leading index is CRIX and for volume weighting LCRIX. While the choice for the best number of constituents is made, their numbers are chosen in steps of 5. It is common in financial industry to construct market indices with a number of constituents which is evenly divisible by 5, see e.g. FTSE (2016), S&P (2014), Deutsche Boerse AG (2013). Therefore this selection is applied for CRIX($k$), $k = 5, 10, 15, \ldots$ with $k_1 = 5$. Since the global minimum for the selection criterion may involve many index constituents, but a sparse index is the goal, the search for the optimal model terminates at level $j$ whenever

$$\text{AIC}\{\hat{\varepsilon}(k_j, \beta), k_j - 5\} < \text{AIC}\{\hat{\varepsilon}(k_{j-1}, \beta), k_{j-1} - 5\} \tag{25}$$

and $k_{j-1}$ index constituents are chosen. Therefore merely a local optimum will be achieved in most of the cases for $\Theta = \Theta_{AIC}$, in (13). But the choice is still asymptotically optimal by defining $\Theta = \{\Theta_{AIC} | k_i \leq k_j \forall i\}$. In Section 6 it will be shown that the performance of the index is already very good.

2. ECRIX/LECRIX: The second constructed index is called Exact CRIX (ECRIX) and Liquidity ECRIX respectively. It follows the above rules too. But the number of its constituents is chosen in steps of 1. Therefore the set of models contains CRIX($k$), $k = 1, 2, 3, \ldots$ with $k_1 = 1$ and stops when

$$\text{AIC}\{\hat{\varepsilon}(k_j, \beta), k_j - 1\} < \text{AIC}\{\hat{\varepsilon}(k_{j-1}, \beta), k_{j-1} - 1\}. \tag{26}$$

3. EFCRIX/LEFCRIX: Since the decision procedures for CRIX and ECRIX terminate when the AIC rises for the first time, Exact Full CRIX and Liquidity EFCRIX will be constructed to visualize whether the decision procedure works fine for the covered indices. The intention is to have an index which may approach the TMI but only in case even small assets help improve the view on the total market, a benchmark for the benchmarks. It will be derived with the AIC procedure, compare Section 3. For $k = 1, 2, 3, \ldots$ with $k_1 = 1$ the decision rule is based on

$$\min_{k_j, \beta} \text{AIC}\{\hat{\varepsilon}(k_j, \beta), k_j - 1\} \tag{27}$$

for $\Theta = \Theta_{AIC}$, in (13). This index computes the AIC for every possible number of constituents and the number is chosen where the AIC becomes minimal.

## 6. Performance analysis

The indices CRIX, ECRIX, EFCRIX with market cap weighting and LCRIX, LECRIX, LEFCRIX with volume weighting have been proposed to give insight into the CC market. Our RDC CC database covers data for over 1000 CCs, kindly provided by CoinGecko. The data used for the analysis cover daily closing data for prices, market volume and market capitalization in USD for each CC in the time period from 2014-04-01 to 2017-03-25. Crypto exchanges are open on the weekends, therefore data for weekend closing prices exist. Since CC exchanges do not finish trading after a certain time point every day, a time point which serves as a closing time has to be defined. CoinGecko used 12 am UTC time zone. One should note that missing data are observed in the dataset, therefore the last rules from Section 4 will come into play.

Fig. 2 shows the performance of CRIX, and Fig. 6 the differences between CRIX and both ECRIX and EFCRIX. For the purpose of comparison, the indices were recalibrated on the recalculation dates since the index constituents change then. We do not provide each index plot individually since they perform almost equally. However, the AIC method gave very different numbers of constituents for the corresponding indices. The numbers of constituents are given in Table 4. For comparison, the number of constituents under the other discussed model selection criteria are provided too. The variance of $C_p$ was derived with a GARCH(1,1) model, Bollerslev (1986). The corresponding information for ECRIX and EFCRIX are given in the same Table 4. Interestingly the methodology of EFCRIX causes its number of constituents to reach a relatively stable value for each period. ECRIX has mostly much fewer constituents than CRIX and EFCRIX due to the fact that this index just runs until a local optimum. Comparing the number of constituents for CRIX derived with AIC against the other criteria, one sees that GC, GFC and SH tend to choose more or the same number of constituents than AIC. Also all three criteria suggest the same result. $C_p$ stops at the initial value for CRIX, ECRIX and EFCRIX. For CRIX, ECRIX and EFCRIX, AIC mostly chooses less constituents compared to all other criteria, except $C_p$ which terminates very early. For LCRIX, LECRIX and LEFCRIX mostly less constituents were chosen than for CRIX, ECRIX and EFCRIX, compare Table 5. Note that the AIC gave the sparsest result again.
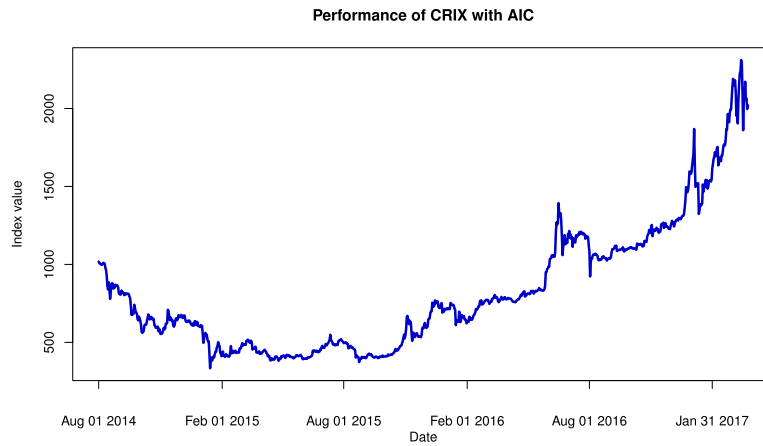
**Performance of CRIX with AIC**



Fig. 2. Performance of CRIX. CRIXindex CRIXcode.

**Table 2**
Comparison of CRIX, ECRIX, EFCRIX, derived under different penalizations, against TMI under mean of monthly Mean Squared Error, compared with btc.

|        | AIC     | GC      | GFC     | SH      | Cp       | FPE    |
|--------|---------|---------|---------|---------|----------|--------|
| CRIX   | 0.4769  | 0.4883  | 0.3755  | 0.3598  | 1.9844   | 0.0042 |
| ECRIX  | 11.0988 | 10.3673 | 10.3673 | 10.4667 | 79.3979  | 0.0048 |
| EFCRIX | 3.1394  | 0.0116  | 0.0049  | 0.0049  | 79.3979  | 0.0048 |
| LCRIX  | 0.6417  | 0.1497  | 0.1217  | 0.1211  | 0.6638   | 0.0049 |
| LECRIX | 22.8782 | 16.7187 | 16.7187 | 16.7187 | 125.0620 | 0.0047 |
| LEFCRIX| 7.9158  | 0.0645  | 0.0126  | 0.0126  | 125.0620 | 0.0047 |
| btc    | 79.3979 | 79.3979 | 79.3979 | 79.3979 | 79.3979  | 79.3979|

**Table 3**
Comparison of CRIX, ECRIX, EFCRIX, derived under different penalizations, against TMI under mean of monthly Mean Directional Accuracy, compared with btc.

|        | AIC    | GC     | GFC    | SH     | Cp     | FPE    |
|--------|--------|--------|--------|--------|--------|--------|
| CRIX   | 0.9896 | 0.9908 | 0.9918 | 0.9928 | 0.9835 | 1.0000 |
| ECRIX  | 0.9576 | 0.9586 | 0.9586 | 0.9586 | 0.9133 | 1.0000 |
| EFCRIX | 0.9794 | 0.9990 | 1.0000 | 1.0000 | 0.9133 | 1.0000 |
| LCRIX  | 0.9928 | 0.9949 | 0.9959 | 0.9959 | 0.9917 | 1.0000 |
| LECRIX | 0.9692 | 0.9700 | 0.9700 | 0.9700 | 0.9501 | 1.0000 |
| LEFCRIX| 0.9855 | 0.9979 | 1.0000 | 1.0000 | 0.9501 | 1.0000 |
| btc    | 0.9133 | 0.9133 | 0.9133 | 0.9133 | 0.9133 | 0.9133 |

The indices optimized until a local optimum are expected to perform less optimal than the globally optimized ones against the TMI/LTMI. Tables 2 and 3 give the mean over monthly Mean Squared Error (MSE) and Mean Directional Accuracy (MDA), defined as

$$\text{MSE}\{\text{CRIX}(k)\} = \frac{1}{t_l^+ - t_l^-} \sum_{t=t_l^-}^{t_l^+} \{\text{CRIX}(k)_t - \text{TMI}(k_{\max})_t\}^2 \tag{28}$$

$$\text{MDA}\{\text{CRIX}(k)\} = \frac{1}{t_l^+ - t_l^-} \sum_{t=t_l^-}^{t_l^+} \mathbf{I}[\text{sign}\{\text{TMI}(k_{\max})_t - \text{TMI}(k_{\max})_{t-1}\}$$
$$= \text{sign}\{\text{CRIX}(k)_t - \text{CRIX}(k)_{t-1}\}] \tag{29}$$

where $t_l^-$ and $t_l^+$ are the beginning and end of the month respectively, $\mathbf{I}(\cdot)$ is the indicator function and sign$(\cdot)$ gives the sign of the respective equation. Apparently CRIX performs best, which can be explained due to its larger number of index constituents. The CRIX, ECRIX and EFCRIX are close in terms of the MDA but the MSE is much better for CRIX. Comparing all the model selection criteria, FPE has the best performance in terms of MSE and MDA, due to choosing high numbers of constituents. The trading volume weighted indices are close in terms of MSE and MDA to their market weighted corresponding indices. At the same time the number of constituents are mostly sparser for the volume weighted ones.

CRIX was constructed with steps of five which is common in practice and performed best under AIC. For this case the number of constituents was the most stable, while achieving the best performance for MSE and MDA. Additionally, the analysis showed that it is indeed unnecessary from a practical viewpoint to choose the global optimal AIC under steps of 1. Even a local optimum and a much
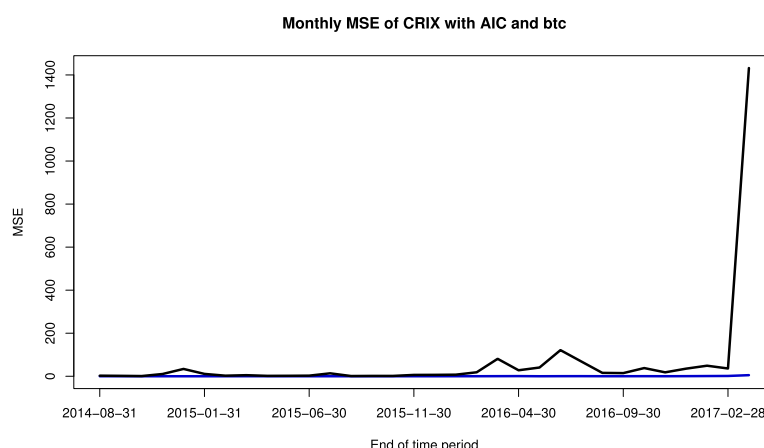
**Fig. 3.** Performance of CRIX compared to BTC.

**Table 4**
Comparison of AIC, GC, GFC, SH, Cp and the FPE method for the selection of the number of index constituents for the CRIX, ECRIX and EFCRIX in the 11 periods.

| | CRIX | | | | | | ECRIX | | | | | | EFCRIX | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | AIC | GC | GFC | SH | Cp | FPE | AIC | GC | GFC | SH | Cp | FPE | AIC | GC | GFC | SH | Cp | FPE | max |
| 1 | 5 | 10 | 10 | 10 | 10 | 35 | 2 | 2 | 2 | 3 | 1 | 36 | 2 | 7 | 30 | 30 | 1 | 36 | 36 |
| 2 | 10 | 15 | 15 | 15 | 5 | 100 | 3 | 3 | 3 | 3 | 1 | 93 | 3 | 94 | 93 | 93 | 1 | 93 | 113 |
| 3 | 5 | 10 | 35 | 35 | 5 | 100 | 5 | 5 | 5 | 5 | 1 | 93 | 5 | 94 | 93 | 93 | 1 | 93 | 158 |
| 4 | 10 | 10 | 10 | 40 | 5 | 95 | 3 | 3 | 3 | 3 | 1 | 90 | 3 | 91 | 90 | 90 | 1 | 90 | 182 |
| 5 | 10 | 20 | 20 | 20 | 5 | 100 | 2 | 4 | 4 | 4 | 1 | 93 | 12 | 94 | 93 | 93 | 1 | 93 | 169 |
| 6 | 10 | 10 | 20 | 20 | 5 | 100 | 2 | 2 | 2 | 2 | 1 | 93 | 2 | 94 | 93 | 93 | 1 | 93 | 171 |
| 7 | 5 | 20 | 20 | 20 | 5 | 100 | 1 | 1 | 1 | 1 | 1 | 93 | 16 | 94 | 93 | 93 | 1 | 93 | 176 |
| 8 | 15 | 20 | 20 | 20 | 5 | 95 | 3 | 4 | 4 | 4 | 1 | 91 | 3 | 92 | 91 | 91 | 1 | 91 | 140 |
| 9 | 15 | 5 | 5 | 5 | 5 | 100 | 3 | 3 | 3 | 3 | 1 | 93 | 3 | 94 | 93 | 93 | 1 | 93 | 188 |
| 10 | 15 | 15 | 25 | 25 | 5 | 100 | 3 | 5 | 5 | 5 | 1 | 93 | 3 | 94 | 93 | 93 | 1 | 93 | 207 |
| 11 | 10 | 35 | 45 | 45 | 5 | 100 | 2 | 2 | 2 | 2 | 1 | 93 | 4 | 94 | 93 | 93 | 1 | 93 | 221 |

more stable number of constituents is able to mimic the market movements very well in terms of the MDA and MSE. Furthermore, even for ECRIX there was more than one constituent selected most of the time. This shows that Bitcoin, which currently clearly dominates the market in terms of market capitalization and trading volume, does not account for all the variance in the market. Other CCs are important for the market movements too.

Depending on the theoretical and empirical analysis, we decided to continue with the AIC. From the theoretical viewpoint, the AIC uses the most information about the data, since it relies on the density. From the empirical analysis, the AIC chooses much less constituents than GC, GFC, SH and FPE, while its performance in terms of MSE and MDA is close to the three outlined criteria. The better performance was achieved due to overparametrization of the index by GC, GFC, SH and FPE. Therefore, CRIX will be derived with the AIC criterion.

Comparing CRIX with the development of BTC, it tracks the market development better over time. Fig. 3 shows the monthly MSE of CRIX with AIC and BTC. In 2016 CRIX tracked the market development much better than BTC, and in the beginning of 2017 even better due to the huge impact of the price gain of altcoins like Ethereum, Ripple and Dash. Their performance is visualized in Fig. 4, clearly showing the better performance of CRIX in this time period, driven by price gains in altcoins. Due to the log scale and the high gains of altcoins, the difference between CRIX and BTC appears little, while in fact being considerable. Fig. 4(b) shows the difference in the log returns of CRIX and BTC. One sees differences in their return series, which are particularly strong beginning of 2016 and in March 2017. Comparing the performance of CRIX and LCRIX against BTC, one observes an increasing spread between the indices, Fig. 5. It indicates a lower weight of BTC in LCRIX, thus tackling the issue of dominance of BTC in CRIX by liquidity weighting. Having a look at the actual differences in the log return series compared to CRIX, Fig. 5(b), stronger spikes are observed, thus showing the difference in the performance from CRIX and LCRIX driven by the stronger weights on altcoins in LCRIX. Table 8 shows the actual weights given to BTC and altcoins in the respective indices. In the liquidity indices altcoins frequently receive a higher weight compared to the respective indices based on market capitalization weighting. Once the altcoins received even 52% of the weights in LCRIX. The results show the market focus in terms of trading is stronger for altcoins than their market capitalization suggests, thus an index accounting for this is called for, LCRIX. Simultaneously the weighting scheme tackles the dominance of BTC in a market capitalization index.
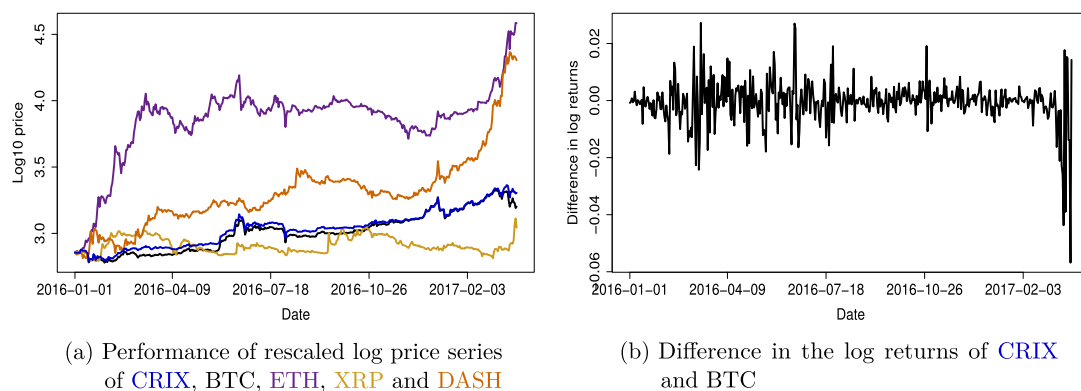
(a) Performance of rescaled log price series of CRIX, BTC, ETH, XRP and DASH

(b) Difference in the log returns of CRIX and BTC

**Fig. 4.** Comparison of performance of CRIX, BTC, ETH, XRP and DASH.



(a) Performance of index series of CRIX, BTC and LCRIX

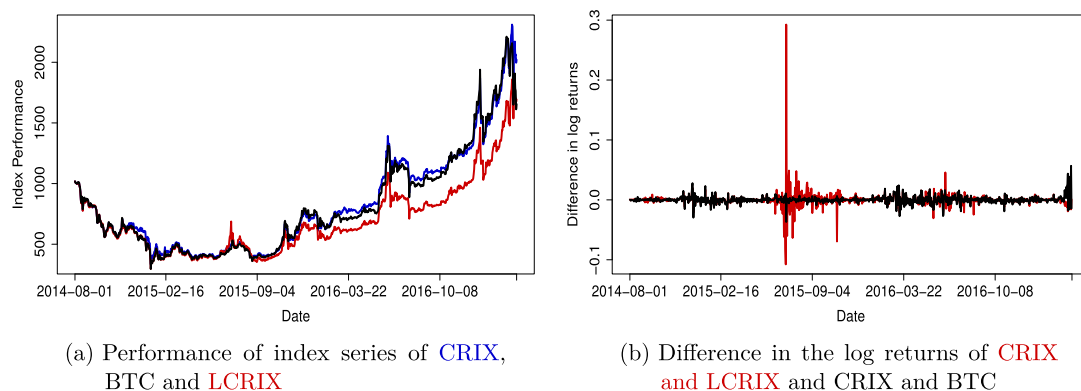(b) Difference in the log returns of CRIX and LCRIX and CRIX and BTC

**Fig. 5.** Comparison of performance of CRIX, BTC and LCRIX.

**Table 5**
Comparison of AIC, GC, GFC, SH, Cp and the FPE method for the selection of the number of index constituents for the LCRIX, LECRIX and LEFCRIX in the 11 periods.

| | LCRIX | | | | | | LECRIX | | | | | | LEFCRIX | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | AIC | GC | GFC | SH | Cp | FPE | AIC | GC | GFC | SH | Cp | FPE | AIC | GC | GFC | SH | Cp | FPE | max |
| 1 | 5 | 10 | 15 | 15 | 5 | 35 | 2 | 3 | 3 | 3 | 1 | 36 | 2 | 6 | 16 | 16 | 1 | 36 | 36 |
| 2 | 5 | 10 | 10 | 10 | 5 | 100 | 2 | 4 | 4 | 4 | 1 | 93 | 2 | 94 | 93 | 93 | 1 | 93 | 113 |
| 3 | 5 | 5 | 20 | 20 | 5 | 100 | 3 | 4 | 4 | 4 | 1 | 93 | 6 | 94 | 93 | 93 | 1 | 93 | 158 |
| 4 | 15 | 20 | 20 | 30 | 5 | 95 | 3 | 2 | 2 | 2 | 1 | 90 | 3 | 91 | 90 | 90 | 1 | 90 | 182 |
| 5 | 5 | 5 | 5 | 5 | 5 | 100 | 1 | 1 | 1 | 1 | 1 | 93 | 93 | 94 | 93 | 93 | 1 | 93 | 169 |
| 6 | 5 | 5 | 5 | 5 | 5 | 100 | 2 | 2 | 2 | 2 | 1 | 93 | 9 | 94 | 93 | 93 | 1 | 93 | 171 |
| 7 | 10 | 25 | 30 | 35 | 5 | 100 | 1 | 2 | 2 | 2 | 1 | 93 | 1 | 94 | 93 | 93 | 1 | 93 | 176 |
| 8 | 10 | 20 | 35 | 35 | 5 | 95 | 1 | 1 | 1 | 1 | 1 | 91 | 3 | 92 | 91 | 91 | 1 | 91 | 140 |
| 9 | 5 | 10 | 10 | 10 | 5 | 100 | 2 | 2 | 2 | 2 | 1 | 93 | 2 | 94 | 93 | 93 | 1 | 93 | 188 |
| 10 | 10 | 10 | 10 | 10 | 5 | 100 | 1 | 1 | 1 | 1 | 1 | 93 | 1 | 94 | 93 | 93 | 1 | 93 | 207 |
| 11 | 5 | 15 | 15 | 15 | 5 | 100 | 2 | 3 | 3 | 3 | 1 | 93 | 2 | 94 | 93 | 93 | 1 | 93 | 221 |

## 7. Application to the German stock market

The CRIX methodology was derived with the idea of finding a method which allows mimicking young and fast changing markets appropriately. But well known major markets usually change their structure too. So the proposed methodology is tested on the German stock market, which has four major indices: DAX, MDAX, SDAX and TecDAX. The DAX is used to determine the overall market direction, Janßen and Rudolph (1992). Since it is chosen from the so called prime segment, it has some prior restrictions. It is interesting to see whether our methodology yields the DAX as an adequate benchmark for the total market. Since the indices are derived with market cap weighting scheme, only this methodology is tested. Following Definition 4, all available stocks are defined as the TMI and our new method is applied to find an appropriate index. Again, the 7-step method from Section 3 was applied to find the number of constituents, but it starts at 30 members to check if more constituents are necessary. The method for the identification of $k$ and the reallocation of the included assets is performed quarterly, like DAX. To be in line with the DAX reallocation dates, the
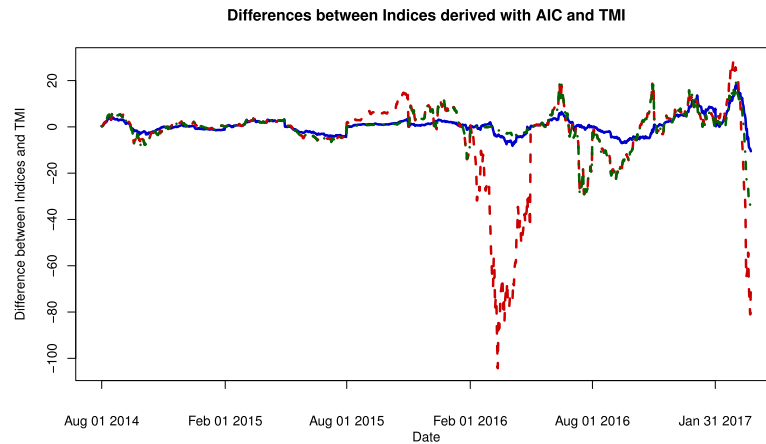
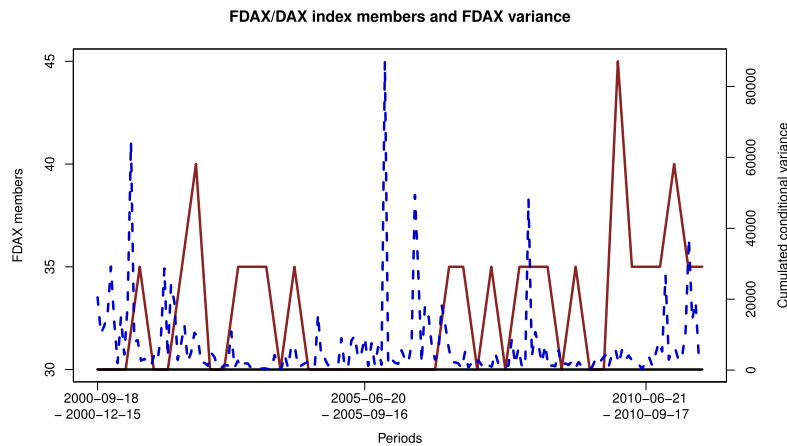**Fig. 6.** Realized difference between TMI and CRIX (solid), ECRIX (dashed), EFCRIX (dotdashed). CRIXfamdiff CRIXcode.



**Fig. 7.** Number of constituents of FDAX (solid), DAX (horizontal solid) and cumulated monthly variance of FDAX (dashed). CRIXdaxmembersvar CRIXcode.

index calculation will start after the third Friday of September and the reallocation dates are the third Fridays of December, March, June and September, see Deutsche Boerse AG (2013).

The data were fetched from Datastream in the period 2000-06-16 until 2015-12-18. All stocks which are German companies and are traded on XETRA are chosen. Any time series for which Datastream reported an error either for the price or market capitalization data was excluded from the analysis. The index, computed with the new methodology, is called Flexible DAX (FDAX). One should note that the analysis starts three months after the starting point of the dataset due to the initialization period of FDAX.

Fig. 7 shows the number of members of FDAX and DAX in the respective periods. Most of the time, the number of index constituents for FDAX is higher than the 30 members of DAX. Just around 2004–2005 is the $k$ more frequently 30. Especially while the turmoil of the financial markets, starting from 2008/2009, is the number of index constituents much higher. One might hint that a higher reported variability in one period should cause an increase in $k$ in the next period, since it was shown that the selection method depends on the variance, see Appendix A. Fig. 7 shows that this idea can partially be supported. The derivation of the conditional variance was performed with a GARCH(1,1) model, Bollerslev (1986), and the daily results were summed up. Obviously, in the extreme cases increases the $k$ in the next period, see 2001, 2002, 2006 and 2011.

The computation of the MSE and MDA, see Table 6, shows that FDAX is a more accurate benchmark for the total market as DAX. Since Janßen and Rudolph (1992) state that DAX may be used to analyze the movements of the total market, an MDA of 92% is indeed good. But FDAX mimics the market even better, with an MDA of 96%. Also the MSE for FDAX is much lower than the one of DAX. Therefore the methodology fulfilled its goal to find a sparse, investable and accurate benchmark, depending on the MDA.

## 8. Application to Mexican stock market

The Mexican stock market is represented by the IPC35, MEXBOL (2013). One of its rules is a readjustment of the weights to lower the effect of dominant stocks. In the CC market BTC is such a dominant asset. The CRIX methodology could help to circumvent arbitrary rules and develop an index to represent the market accurately.

**Table 6**

Comparison of DAX with CRIX methodology (FDAX) and rescaled DAX against TMI.

|              | MSE   | MDA  |
|--------------|-------|------|
| FDAX vs. TMI | 6.36  | 0.96 |
| DAX vs. TMI  | 51.02 | 0.92 |



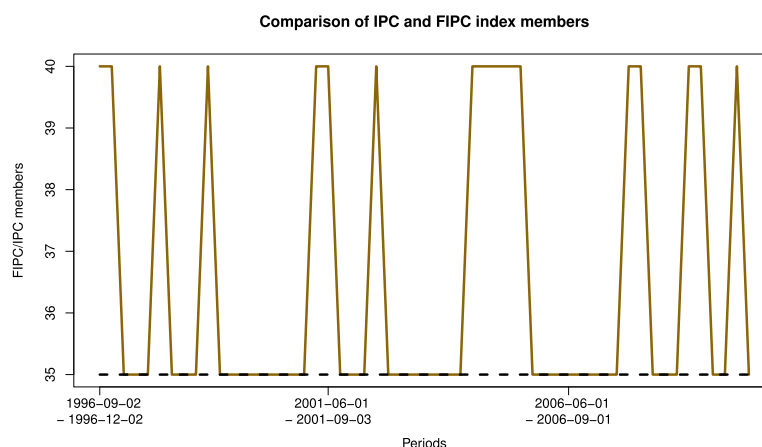**Fig. 8.** Number of constituents of FIPC (solid) and IPC (dashed) in the respective periods. CRIXipcmembers CRIXcode.

**Table 7**

Comparison of IPC with CRIX methodology (FIPC) and rescaled IPC against TMI.

|              | MSE     | MDA  |
|--------------|---------|------|
| FIPC vs. TMI | 24.97   | 0.97 |
| IPC vs. TMI  | 4743.50 | 0.91 |

The data were fetched from Datastream for the period 1996-06-01 until 2015-05-29 and cover all Mexican companies listed in Datastream. The specifications of the methodology are the same as for the German stock market except for the recalculation date. In line with the methodology of the IPC35, the index is recalculated with the closing data of the last business days of August, November, February and May, therefore the recalculated index starts on the first business days of September, December, March and June. The TMI will be all fetched companies. The choice of $k$ starts with 35 since this is the amount of constituents of IPC.

Again, the CRIX methodology works well. The MSE is very low compared to the one for the IPC35 and the MDA gives a much better performance too, see Table 7. We can conclude that the methodology helped to circumvent the usage of arbitrary rules for the weights in the rules of the indices and enhances at the same time the performance of the market index. Fig. 8 shows the number of index members of the FIPC compared to the IPC. Obviously, the methodology also suggests using more than 35 index members half of the time which is the number of members of the IPC.

## 9. Conclusion

The movements of CCs are very different from each other, Elendner et al. (2017). So studying the entire market of CCs requires an instrument which adequately captures and displays the market movements, an index. But index construction for CCs requires a new methodology to find the right number of index members. Innovative markets, like the one for CCs, change their structure frequently. The proposed methods were applied to oracle a new family of indices, which are displayed and updated on a daily basis. The performance of the new indices were studied and it was shown that the dynamic AIC based methodology results in indices with stable properties. The results show that a market like the CC market – momentarily dominated by Bitcoin – still needs a representative index since Bitcoin does not account for all the variance in the market. The diversified nature of the CC market makes the inclusion of altcoins in the index product critical to improve tracking performance. We have shown that assigning optimal weights to altcoins helps to reduce the tracking errors of a CC portfolio, despite the fact that their market cap is much smaller relative to Bitcoin.

Besides the classical market capitalization weighting, a volume weighting scheme was proposed. The corresponding indices are sparser in terms of constituents while having a comparable performance, which gives support to this weighting scheme under the goals of the study. The AIC based method was also applied to the German stock market. The results yield a more accurate benchmark in terms of MDA. In applying the CRIX methodology to the Mexican stock market, which is dominated by Telmex, one finds high accuracy of it in terms of MSE and MDA.

We conclude, that the CRIX technology enhances the construction of an index if the goal is to find a sparse, investable and accurate benchmark.

**Table 8**

Average monthly weights of BTC and altcoins in the respective periods in the 6 indices.

| Periods | CRIX | | | LCRIX | | | ECRIX | | | LECRIX | | | EFCRIX | | | LEFCRIX | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | k | BTC | altcoins | k | BTC | altcoins | k | BTC | altcoins | k | BTC | altcoins | k | BTC | altcoins | k | BTC | altcoins |
| 2014/08 | 5 | 0.96 | 0.04 | 5 | 0.96 | 0.04 | 2 | 0.98 | 0.02 | 2 | 0.97 | 0.03 | 2 | 0.98 | 0.02 | 2 | 0.97 | 0.03 |
| 2014/09 | 5 | 0.94 | 0.06 | 5 | 0.82 | 0.18 | 2 | 0.97 | 0.03 | 2 | 0.86 | 0.14 | 2 | 0.97 | 0.03 | 2 | 0.86 | 0.14 |
| 2014/10 | 5 | 0.93 | 0.07 | 5 | 0.86 | 0.14 | 2 | 0.97 | 0.03 | 2 | 0.91 | 0.09 | 2 | 0.97 | 0.03 | 2 | 0.91 | 0.09 |
| 2014/11 | 10 | 0.92 | 0.08 | 5 | 0.95 | 0.05 | 3 | 0.94 | 0.06 | 2 | 0.99 | 0.01 | 3 | 0.94 | 0.06 | 2 | 0.99 | 0.01 |
| 2014/12 | 10 | 0.85 | 0.15 | 5 | 0.95 | 0.05 | 3 | 0.86 | 0.14 | 2 | 0.97 | 0.03 | 3 | 0.86 | 0.14 | 2 | 0.97 | 0.03 |
| 2015/01 | 10 | 0.82 | 0.18 | 5 | 0.93 | 0.07 | 3 | 0.85 | 0.15 | 2 | 0.98 | 0.02 | 3 | 0.85 | 0.15 | 2 | 0.98 | 0.02 |
| 2015/02 | 5 | 0.86 | 0.14 | 5 | 0.90 | 0.10 | 5 | 0.86 | 0.14 | 3 | 0.91 | 0.09 | 5 | 0.86 | 0.14 | 6 | 0.90 | 0.10 |
| 2015/03 | 5 | 0.90 | 0.10 | 5 | 0.96 | 0.04 | 5 | 0.90 | 0.10 | 3 | 0.96 | 0.04 | 5 | 0.90 | 0.10 | 6 | 0.96 | 0.04 |
| 2015/04 | 5 | 0.90 | 0.10 | 5 | 0.94 | 0.06 | 5 | 0.90 | 0.10 | 3 | 0.95 | 0.05 | 5 | 0.90 | 0.10 | 6 | 0.94 | 0.06 |
| 2015/05 | 10 | 0.90 | 0.10 | 15 | 0.96 | 0.04 | 3 | 0.92 | 0.08 | 3 | 0.96 | 0.04 | 3 | 0.92 | 0.08 | 3 | 0.96 | 0.04 |
| 2015/06 | 10 | 0.87 | 0.13 | 15 | 0.86 | 0.14 | 3 | 0.90 | 0.10 | 3 | 0.88 | 0.12 | 3 | 0.90 | 0.10 | 3 | 0.88 | 0.12 |
| 2015/07 | 10 | 0.88 | 0.12 | 15 | 0.48 | 0.52 | 3 | 0.90 | 0.10 | 3 | 0.48 | 0.52 | 3 | 0.90 | 0.10 | 3 | 0.48 | 0.52 |
| 2015/08 | 10 | 0.88 | 0.12 | 5 | 0.59 | 0.41 | 2 | 0.93 | 0.07 | 1 | 1.00 | 0.00 | 12 | 0.88 | 0.12 | 93 | 0.58 | 0.42 |
| 2015/09 | 10 | 0.89 | 0.11 | 5 | 0.53 | 0.47 | 2 | 0.93 | 0.07 | 1 | 1.00 | 0.00 | 12 | 0.88 | 0.12 | 93 | 0.52 | 0.48 |
| 2015/10 | 10 | 0.90 | 0.10 | 5 | 0.59 | 0.41 | 2 | 0.96 | 0.04 | 1 | 1.00 | 0.00 | 12 | 0.90 | 0.10 | 93 | 0.58 | 0.42 |
| 2015/11 | 10 | 0.92 | 0.08 | 5 | 0.82 | 0.18 | 2 | 0.97 | 0.03 | 2 | 0.83 | 0.17 | 2 | 0.97 | 0.03 | 9 | 0.82 | 0.18 |
| 2015/12 | 10 | 0.93 | 0.07 | 5 | 0.84 | 0.16 | 2 | 0.98 | 0.02 | 2 | 0.84 | 0.16 | 2 | 0.98 | 0.02 | 9 | 0.84 | 0.16 |
| 2016/01 | 10 | 0.92 | 0.08 | 5 | 0.87 | 0.13 | 2 | 0.97 | 0.03 | 2 | 0.87 | 0.13 | 2 | 0.97 | 0.03 | 9 | 0.87 | 0.13 |
| 2016/02 | 5 | 0.89 | 0.11 | 10 | 0.90 | 0.10 | 1 | 1.00 | 0.00 | 1 | 1.00 | 0.00 | 16 | 0.87 | 0.13 | 1 | 1.00 | 0.00 |
| 2016/03 | 5 | 0.83 | 0.17 | 10 | 0.86 | 0.14 | 1 | 1.00 | 0.00 | 1 | 1.00 | 0.00 | 16 | 0.81 | 0.19 | 1 | 1.00 | 0.00 |
| 2016/04 | 5 | 0.85 | 0.15 | 10 | 0.93 | 0.07 | 1 | 1.00 | 0.00 | 1 | 1.00 | 0.00 | 16 | 0.84 | 0.16 | 1 | 1.00 | 0.00 |
| 2016/05 | 15 | 0.83 | 0.17 | 10 | 0.75 | 0.25 | 3 | 0.87 | 0.13 | 1 | 1.00 | 0.00 | 3 | 0.87 | 0.13 | 3 | 0.75 | 0.25 |
| 2016/06 | 15 | 0.85 | 0.15 | 10 | 0.65 | 0.35 | 3 | 0.88 | 0.12 | 1 | 1.00 | 0.00 | 3 | 0.88 | 0.12 | 3 | 0.65 | 0.35 |
| 2016/07 | 15 | 0.84 | 0.16 | 10 | 0.71 | 0.29 | 3 | 0.90 | 0.10 | 1 | 1.00 | 0.00 | 3 | 0.90 | 0.10 | 3 | 0.71 | 0.29 |
| 2016/08 | 15 | 0.83 | 0.17 | 5 | 0.70 | 0.30 | 3 | 0.90 | 0.10 | 2 | 0.72 | 0.28 | 3 | 0.90 | 0.10 | 2 | 0.72 | 0.28 |
| 2016/09 | 15 | 0.82 | 0.18 | 5 | 0.73 | 0.27 | 3 | 0.88 | 0.12 | 2 | 0.77 | 0.23 | 3 | 0.88 | 0.12 | 2 | 0.77 | 0.23 |
| 2016/10 | 15 | 0.84 | 0.16 | 5 | 0.84 | 0.16 | 3 | 0.89 | 0.11 | 2 | 0.85 | 0.15 | 3 | 0.89 | 0.11 | 2 | 0.85 | 0.15 |
| 2016/11 | 15 | 0.87 | 0.13 | 10 | 0.94 | 0.06 | 3 | 0.91 | 0.09 | 1 | 1.00 | 0.00 | 3 | 0.91 | 0.09 | 1 | 1.00 | 0.00 |
| 2016/12 | 15 | 0.89 | 0.11 | 10 | 0.94 | 0.06 | 3 | 0.93 | 0.07 | 1 | 1.00 | 0.00 | 3 | 0.93 | 0.07 | 1 | 1.00 | 0.00 |
| 2017/01 | 15 | 0.88 | 0.12 | 10 | 0.92 | 0.08 | 3 | 0.93 | 0.07 | 1 | 1.00 | 0.00 | 3 | 0.93 | 0.07 | 1 | 1.00 | 0.00 |
| 2017/02 | 10 | 0.89 | 0.11 | 5 | 0.93 | 0.07 | 2 | 0.94 | 0.06 | 2 | 0.94 | 0.06 | 4 | 0.92 | 0.08 | 2 | 0.94 | 0.06 |
| 2017/03 | 10 | 0.81 | 0.19 | 5 | 0.74 | 0.26 | 2 | 0.87 | 0.13 | 2 | 0.82 | 0.18 | 4 | 0.84 | 0.16 | 2 | 0.82 | 0.18 |

## Appendix A

### A.1. Proof of Theorem 1

**Proof.** Assume normally distributed error terms, (10) and (22): $\varepsilon(k,\beta) \sim \mathrm{N}\{0, \sigma(k,\beta)^2\}$, $\hat{\varepsilon}(k,\beta) \sim \mathrm{N}\{0, \hat{\sigma}(k,\beta)^2\}$. Then

$$\log L\{\varepsilon(k,\beta)\} = -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log\sigma(k,\beta)^2 - \frac{1}{2\sigma(k,\beta)^2}\sum_{t=1}^{T}\varepsilon(k,\beta)_t^2. \tag{30}$$

Denote $RSS\{\hat{\varepsilon}(k,\beta)\} = \sum_{t=1}^{T}\hat{\varepsilon}(k,\beta)_t^2$ and $\hat{\sigma}(k,\beta)^2 = T^{-1}RSS\{\hat{\varepsilon}(k,\beta)\}$. Then

$$\log L\{\hat{\varepsilon}(k,\beta)\} = -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log T^{-1}RSS\{\hat{\varepsilon}(k,\beta)\} - \frac{1}{2T^{-1}RSS\{\hat{\varepsilon}(k,\beta)\}}RSS\{\hat{\varepsilon}(k,\beta)\} \tag{31}$$

$$= -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log T^{-1}RSS\{\hat{\varepsilon}(k,\beta)\} - \frac{T}{2} \tag{32}$$

$$= -\frac{T}{2}\log T^{-1}RSS\{\hat{\varepsilon}(k,\beta)\} + C \tag{33}$$

with $C = -\frac{T}{2}\log(2\pi) - \frac{T}{2}$. Since $C$ does not depend on any model parameters, just on the data length $T$, this part of the equation could be omitted.

$$\mathrm{AIC}\{\hat{\varepsilon}(k,\beta), s\} = T\log T^{-1}RSS\{\hat{\varepsilon}(k,\beta)\} + 2 \cdot s \tag{34}$$

$$= T\log\hat{\sigma}(k,\beta)^2 + 2 \cdot s \tag{35}$$

The enhancement in the fit to the Total Market Index (TMI) by adding more constituents, $s$, determines the degree of improvement of the likelihood.

With the linearity property of the expectation operator, assume without loss of generality

$$E\{\varepsilon(k_{max})^{TM}\} = E\{\varepsilon(k, \beta)^{CRIX}\} = 0$$
$$t \in \{1, \dots, T\}$$
$$t_l^- = 0$$
$$s = 1$$

$$\begin{aligned}
\hat{\sigma}(k, \beta) &= \mathrm{Var}\{\hat{\varepsilon}(k, \beta)\} \\
&= \mathrm{Var}\{\varepsilon(k_{max})^{TM} - \varepsilon(k, \beta)^{CRIX}\} \\
&= \sum_{t=1}^{T}\Bigg[\log\Bigg\{\sum_{i=1}^{k_{max}} P_{it}Q_{i,0}(\sum_{i=1}^{k} P_{i,t-1}Q_{i,0} + \beta_1 P_{k+1,t-1}Q_{k+1,0})\Bigg\} \\
&\qquad - \log\Bigg\{\sum_{i=1}^{k_{max}} P_{i,t-1}Q_{i,0}(\sum_{i=1}^{k} P_{i,t}Q_{i,0} + \beta_1 P_{k+1,t}Q_{k+1,0})\Bigg\}\Bigg]^2 \\
&= \sum_{t=1}^{T}\Bigg[\log\Bigg\{\sum_{i=1}^{k_{max}} P_{it}Q_{i,0} \sum_{i=1}^{k} P_{i,t-1}Q_{i,0} + \sum_{i=1}^{k_{max}} P_{it}Q_{i,0}\beta_1 P_{k+1,t-1}Q_{k+1,0}\Bigg\} \\
&\qquad - \log\Bigg\{\sum_{i=1}^{k_{max}} P_{i,t-1}Q_{i,0} \sum_{i=1}^{k} P_{i,t}Q_{i,0} + \sum_{i=1}^{k_{max}} P_{i,t-1}Q_{i,0}\beta_1 P_{k+1,t}Q_{k+1,0}\Bigg\}\Bigg]^2
\end{aligned}$$

Using the relation $\log(a + b) = \log(a) + \log(1 + \frac{b}{a})$, it results:

$$\begin{aligned}
&= \sum_{t=1}^{T}\Bigg[\log\Bigg\{\sum_{i=1}^{k_{max}} P_{it}Q_{i,0} \sum_{i=1}^{k} P_{i,t-1}Q_{i,0}\Bigg\} + \log\Bigg\{1 + \frac{\sum_{i=1}^{k_{max}} P_{it}Q_{i,0}\beta_1 P_{k+1,t-1}Q_{k+1,0}}{\sum_{i=1}^{k_{max}} P_{it}Q_{i,0} \sum_{i=1}^{k} P_{i,t-1}Q_{i,0}}\Bigg\} \\
&\qquad - \log\Bigg\{\sum_{i=1}^{k_{max}} P_{i,t-1}Q_{i,0} \sum_{i=1}^{k} P_{i,t}Q_{i,0}\Bigg\} + \log\Bigg\{1 + \frac{\sum_{i=1}^{k_{max}} P_{i,t-1}Q_{i,0}\beta_1 P_{k+1,t}Q_{k+1,0}}{\sum_{i=1}^{k_{max}} P_{i,t-1}Q_{i,0} \sum_{i=1}^{k} P_{i,t}Q_{i,0}}\Bigg\}\Bigg]^2 \\
&= \sum_{t=1}^{T}\Bigg(\log\Bigg\{\sum_{i=1}^{k_{max}} P_{it}Q_{i,0} \sum_{i=1}^{k} P_{i,t-1}Q_{i,0}\Bigg\} - \log\Bigg\{\sum_{i=1}^{k_{max}} P_{i,t-1}Q_{i,0} \sum_{i=1}^{k} P_{i,t}Q_{i,0}\Bigg\} \\
&\qquad + \Bigg[\log\Bigg\{1 + \frac{\beta_1 P_{k+1,t-1}Q_{k+1,0}}{\sum_{i=1}^{k} P_{i,t-1}Q_{i,0}}\Bigg\} - \log\Bigg\{1 + \frac{\beta_1 P_{k+1,t}Q_{k+1,0}}{\sum_{i=1}^{k} P_{i,t}Q_{i,0}}\Bigg\}\Bigg]\Bigg)^2
\end{aligned}$$
(36)

Solving the derivation and writing the terms which do not depend on $\beta_1$ as $A_t$ and the last part of (36) as $B_t$:

$$\begin{aligned}
\hat{\sigma}(k, \beta) &= \sum_{t=1}^{T} A_t + 2\log\Bigg\{\sum_{i=1}^{k_{max}} P_{it}Q_{i,0} \sum_{i=1}^{k} P_{i,t-1}Q_{i,0}\Bigg\} B_t - 2\log\Bigg\{\sum_{i=1}^{k_{max}} P_{i,t-1}Q_{i,0} \sum_{i=1}^{k} P_{i,t}Q_{i,0}\Bigg\} B_t + B_t^2 \\
&= \sum_{t=1}^{T} A_t + 2B_t\Bigg[\log\Bigg\{\sum_{i=1}^{k_{max}} P_{it}Q_{i,0} \sum_{i=1}^{k} P_{i,t-1}Q_{i,0}\Bigg\} - \log\Bigg\{\sum_{i=1}^{k_{max}} P_{i,t-1}Q_{i,0} \sum_{i=1}^{k} P_{i,t}Q_{i,0}\Bigg\}\Bigg] + B_t^2 \\
&= \sum_{t=1}^{T} A_t + 2B_t\left[\varepsilon(k_{max})^{TM} - \varepsilon(k, 1)^{CRIX}\right] + B_t^2
\end{aligned}$$

Since normally distributed error terms are assumed, note that $\beta_1 = \frac{Cov\{\hat{\varepsilon}(k,1),\varepsilon_{k+1}\}}{Var\{\varepsilon_{k+1}\}}$, where $\varepsilon_{k+1}$ is the log return of $P_{i,t}Q_{i,0}$. The change in the variance will depend on the additional variance which the new constituent can explain, see $\beta_1$. Furthermore, it depends on the value of $P_{k+1,t}Q_{k+1,0}$ relative to $\sum_{i=1}^{k} P_{i,t}Q_{i,0}$, (36), which is the summed market value of the constituents in the index. This infers that constituents with a higher market capitalization are more likely to be part of the index. ∎

This gives support to using the often applied top-down approach, which we use for the construction of CRIX too.

## References

Akaike, H., 1970. Statistical predictor identification. Ann. Inst. Statist. Math. 22 (1), 203–217.

Akaike, H., 1998. Information theory and an extension of the maximum likelihood principle. In: Parzen, E., Tanabe, K., Kitagawa, G. (Eds.), Selected Papers of Hirotugu Akaike. In: Springer Series in Statistics, Springer New York, pp. 199–213.

Arlot, S., Celisse, A., 2010. A survey of cross-validation procedures for model selection. Stat. Surv. 4, 40–79.

Bollerslev, T., 1986. Generalized autoregressive conditional heteroskedasticity. J. Econometrics 31 (3), 307–327.

Bolt, W., Oordt, M.v., 2016. On the Value of Virtual Currencies. SSRN Scholarly Paper ID 2842557. Social Science Research Network, Rochester, NY.

Chen, S., Chen, C.Y.-H., Härdle, W.K., Lee, T.M., Ong, B., 2017. Econometric analysis of a cryptocurrency index for portfolio investment. In: Lee Kuo Chuen, D., Deng, R. (Eds.), Handbook of Digital Finance and Financial Inclusion: Cryptocurrency, FinTech, InsurTech, and Regulation, vol. 1. Elsevier.

Craven, P., Wahba, G., 1978. Smoothing noisy data with spline functions. Numer. Math. 31 (4), 377–403.

CRSP, 2015. CRSP U.S. Equity Indexes Methodology Guide, crsp.com/.

Deutsche Boerse AG, 2013. Guide to the equity Indizes of Deutsche Boerse, www.dax-indices.com.

Devroye, L., Györfi, L., 1985. Nonparametric Density Estimation The L1 View. Wiley.

Droge, B., 1996. Some comments on cross-validation. In: Härdle, W.K., Schimek, M.G. (Eds.), Statistical Theory and Computational Aspects of Smoothing. In: Contributions to Statistics, Physica-Verlag HD, pp. 178–199.

EconoTimes, 2016. Japans cabinet Approves New Bitcoin Regulations, econotimes.com.

Elendner, H., Trimborn, S., Ong, B., Lee, T.M., 2017. The cross-section of crypto-currencies as financial assets. In: Lee Kuo Chuen, D., Deng, R. (Eds.), Handbook of Digital Finance and Financial Inclusion: Cryptocurrency, FinTech, InsurTech, and Regulation, vol. 1. Elsevier.

Epanechnikov, V., 1969. Non-parametric Estimation of a Multivariate Probability Density. Theory Probab. Appl. 14 (1), 153–158.

FTSE, 2016. FTSE UK Index Series, www.ftse.com.

Györfi, L., Härdle, W.K., Sarda, P., Vieu, P., 1989. Nonparametric curve Estimation from Time Series. In: Györfi, L., Härdle, W.K., Sarda, P., Vieu, P. (Eds.), Lecture Notes in Statistics, vol. 60. Springer New York.

Hall, P., 1987. On Kullback-Leibler loss and density estimation. Ann. Statist. 15 (4), 1491–1519.

Härdle, W.K., Müller, M., Sperlich, S., Werwatz, A., 2004. Nonparametric and Semiparametric Models. Springer Science & Business Media.

Härdle, W.K., Trimborn, S., 2015. CRIX or evaluating Blockchain based currencies, Oberwolfach Report No. 42/2015 "The Mathematics and Statistics of Quantitative Risk.

Hayek, F.A., 1990. Denationalization of Money: An Analysis of the Theory and Practice of Concurrent Currencies, third ed. Institute of Economic Affairs, London.

Horton, N.J., Kleinman, K.P., 2007. Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. Amer. Statist. 61 (1), 79–90.

Hurvich, C.M., Tsai, C.-L., 1989. Regression and time series model selection in small samples. Biometrika 76 (2), 297–307.

Janßen, B., Rudolph, B., 1992. Der deutsche Aktienindex DAX. Fritz Knapp Verlag.

Kanazawa, Y., 1993. Hellinger distance and Kullback–Leibler loss for the kernel density estimator. Statist. Probab. Lett. 18 (4), 315–321.

Kawa, L., 2015. Bitcoin Is Officially a Commodity, According to U.S. Regulator, Bloomberg.com.

Kristoufek, L., 2015. What are the main drivers of the Bitcoin price? Evidence from wavelet coherence analysis. PLoS One 10, 1–15.

Mallick, H., Yi, N., 2013. Bayesian methods for High Dimensional Linear Models. J. Biometr. Biostat. 1.

Mallows, C.L., 1973. Some comments on Cp. Technometrics 15 (4), 661–675.

MEXBOL, 2013. Prices and quotations Index (MEXBOL) - Methodology Note, bmv.com.

Nishii, R., 1984. Asymptotic properties of criteria for selection of variables in multiple regression. Ann. Statist. 12 (2), 758–765.

Reid, F., Harrigan, M., 2013. An analysis of anonymity in the bitcoin system. In: Altshuler, Y., Elovici, Y., Cremers, A.B., Aharony, N., Pentland, A. (Eds.), Security and Privacy in Social Networks. Springer New York, pp. 197–223.

Ron, D., Shamir, A., 2013. Quantitative analysis of the full bitcoin transaction graph. In: Sadeghi, A.-R. (Ed.), Financial Cryptography and Data Security. In: Lecture Notes in Computer Science, vol. 7859, Springer Berlin Heidelberg, pp. 6–24.

Sheather, S.J., Jones, M.C., 1991. A reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation. Journal of the Royal Statistical Society. Series B. Methodological 53, 683–690.

Shibata, R., 1981. An optimal Selection of Regression Variables. Biometrika 68 (1), 45–54.

Shibata, R., 1983. Asymptotic mean efficiency of a selection of regression variables. Ann. Inst. Statist. Math. 35 (1), 415–423.

S&P, 2014. Index mathematics - Methodology, us.spindices.com.

S&P, 2015. Dow jones Total Stock Market Indices Methodology, us.spindices.com.

Wand, M.P., Jones, M.C., 1994. Multivariate plug-in bandwidth selection. Comput. Stat. 9 (2), 97–116.

Wang, S., Vergne, J.P., 2017. Buzz factor or innovation potential: What explains cryptocurrencies returns?. PLoS One 12 (1), e0169556.

White, L.H., 2014. The market for Cryptocurrencies. SSRN Scholarly Pape ID 2538290. Social Science Research Network, Rochester, NY.

Wilshire Associates, 2015. Wilshire 5000 Total Market Index Methodology, wilshire.com.

Woodroofe, M., 1982. On model selection and the ARC sine laws. Ann. Statist. 10 (4), 1182–1194.