SFB 649 Discussion Paper 2016-021

CRIX an Index for blockchain based Currencies

Simon Trimborn*
Wolfgang Karl Härdle*

* Humboldt-Universität zu Berlin, Germany

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".

http://sfb649.wiwi.hu-berlin.de
ISSN 1860-5664

SFB 649, Humboldt-Universität zu Berlin
Spandauer Straße 1, D-10178 Berlin

CRIX an Index for blockchain based currencies [1]

Simon Trimborn [2]
Wolfgang Karl Härdle [3]

November 3, 2016

The S&P500 or DAX30 are important benchmarks for the financial industry. These and other indices describe different compositions of certain segments of the financial markets. For currency markets, the IMF offers the index SDR. Prior to the Euro, the ECU existed, which was an index representing the development of European currencies. It is surprising, though, to see that the common index providers have not mapped emerging e-coins into an index yet because with cryptos like Bitcoin, a new kind of asset of great public interest has arisen. Index providers decide on a fixed number of index constituents which will represent the market segment. It is a huge challenge to set this fixed number and develop the rules to find the constituents, especially since markets change and this has to be taken into account. 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. The codes used to obtain the results in this paper are available via `www.quantlet.de`  .

JEL classification: C51, C52, G10

Keywords: Index construction, model selection, AIC, bitcoin, cryptocurrency, CRIX

[1]Financial support from the Deutsche Forschungsgemeinschaft via CRC 649 "Economic Risk" and IRTG 1792 "High Dimensional Non Stationary Time Series", Humboldt-Universität zu Berlin, is gratefully acknowledged.

[2]Humboldt-Universität zu Berlin, C.A.S.E. - Center for Applied Statistics and Economics, Spandauer Str. 1, 10178 Berlin, Germany, tel: +49 (0)30 2093-5728, fax: +49 (0)30 2093-5649, E-Mail: simon.trimborn@wiwi.hu-berlin.de

[3]Humboldt-Universität zu Berlin, C.A.S.E. - Center for Applied Statistics and Economics, Spandauer Str. 1, 10178 Berlin, Germany and SKBI School of Business, Singapore Management University, 50 Stamford Road, Singapore 178899, tel: +49 (0)30 2093-5630, E-Mail: haerdle@hu-berlin.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. The idea of the Nobel Laureate Hayek, see Hayek (1990), of letting companies offer concurrent currencies seemed for a long time scarcely probable, but the invention of the *Blockchain* has made it possible to bring his vision to life. Cryptocurrencies (abbr. cryptos) have surfaced and opened up an angle towards this new level of economic interaction. Since the appearance of bitcoins, several new cryptos have spread through the Web and offered new ways of proliferation. Even states accept them as legal payment method or part of economic interaction. E.g., the USA classifies cryptocurrencies 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.

The technical aspects behind cryptocurrencies have been reviewed by several researchers. For a well written technical survey, we refer to Tschorsch and Scheuermann (2015). The transaction graph of Bitcoin, the Blockchain, has received much attention too, see e.g. Ron and Shamir (2013) and Reid and Harrigan (2013). Even the economics of the Bitcoin has been studied, e.g. Kristoufek (2014). To our knowledge, the development of the entire cryptocurrency market has not been studied so far, only subsamples have been taken into account. Elendner et al. (2016) studied the top 10 cryptocurrencies by market capitalization and found that their returns are weakly correlated which each other. This brings us to the conclusion that Bitcoin, even though it dominates the market in terms of its market capitalization, does not lead the market. The movements of other cryptocurrencies are important too, when one analyzes the market of cryptocurrencies. We contribute to this area of research by designing CRIX, a market index (benchmark) which will enable each interested party to study the performance of the crypto market as a whole or single cryptos. Studying the stochastic dynamics of CRIX will allow to create ETFs or contingent claims.

First, the term benchmark has to be defined:

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

Index providers construct their indices in terms of Definition 1 with a fixed number of constituents, see e.g. FTSE (2016), S&P (2014) and Deutsche Boerse AG (2013). But markets change which should cause the chosen number of index constituents to be altered too. While trying to mimic the movements of an innovative market like the crypto market, one is confronted with a frequently changing market structure. This calls for a dynamic structure of the benchmark, especially for the number of constituents. The StrataQuant index family, see NYSE (2015), for example alters the number of constituents in each sector index dependent on their affiliation with a certain sector and membership in the Russell1000 index. But the benchmark for the crypto market would not have a parent index

since it is meant to be the leading index. Therefore a different approach is necessary that enables to react to changes in the market structure. A dynamic methodology guaranteeing the diversity of an index at any time is to be constructed. Furthermore, the benchmark is meant to be investable. Regarding the portfolio choice, we define the following selection definition:

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

This definition corresponds to Occam's Razor. Following definitions 1 and 2, for the crypto market CRIX: a CRyptocurrency IndeX, hu.berlin/crix has been established. To compute CRIX, we evaluate the differences in the log returns of the market against a selection of possible benchmarks. We figure out, that the AIC works well to evaluate the differences. It penalizes the index for the number of constituents, so definitions 1 and 2 are met. For the calculation of the respective likelihoods, a non-parametric approach using the Epanechnikov (1969) kernel is applied. We proof the impact of the value of an asset in the market on the AIC method, thus we are applying a top-down approach to select the assets for the benchmarks to choose from. 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. (2016), or study the crypto 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 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. 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) can not 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, hu.berlin/crix, 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, β_{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 β_{i,t_l^-} were updated. In the classical setting, β_{i,t_l^-} is defined to be $\beta_{i,t_l^-} = 1$ for all i and l. Anyhow, some indices use β_{i,t_l^-} to achieve maximal 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

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

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 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. 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 praxis has shown that smaller indices like DAX30 and S&P500 receive more attention in evaluating the movements of their corresponding markets. 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. Figure 1 shows that this is indeed a problem in the crypto-currencies market. Some cryptos have a fairly high market capitalization while their respective trading volume is very low. An asset which is not frequently traded can not add enough information to a

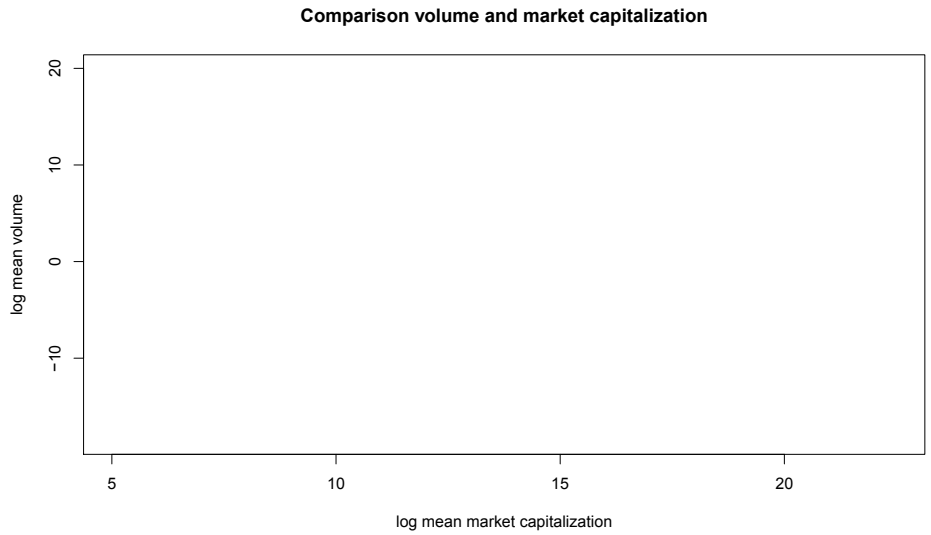

Figure 1: Comparison of the log mean trading volume and log mean market capitalization, both measured in USD, for all cryptos in the dataset over the time period 20140401 - 20160807

VolMarketCapComparison

market index to display market changes and is difficult to trade for an investor.

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 crypto benchmark with sparse number of constituents is asked for. Since the crypto market shows a frequently changing market structure with a huge number of illiquid cryptos, we apply a time varying index selection structure.

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. Since CRIX will be a benchmark for the crypto market, the dimension and evaluation of the market has to be defined:

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

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

$$\mathrm{TM}(K)_t = \frac{\sum_{i=1}^{K} P_{it} Q_{i,t_l^-}}{Divisor(K)_{t_l^-}} \tag{5}$$

with K the number of all cryptos in the crypto universe. Note that no adjustment factor is used for $\mathrm{TM}(K)_t$. Further define the log returns:

$$\varepsilon(K)_t^{TM} = \log\{\mathrm{TM}(K)_t\} - \log\{\mathrm{TM}(K)_{t-1}\} \tag{6}$$
$$\varepsilon(k,\beta)_t^{CRIX} = \log\{\mathrm{CRIX}(k,\beta)_t\} - \log\{\mathrm{CRIX}(k,\beta)_{t-1}\}, \tag{7}$$

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

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

$$\min_{k,\beta} \|\varepsilon(k,\beta)\|^2 = \|\varepsilon(K)^{TM} - \varepsilon(k,\beta)^{CRIX}\|^2, \tag{8}$$

is achieved. We chose a squared loss function in (8), since it heavily penalizes deviations. The expected squared loss is defined as

$$\mathsf{E}(\|\varepsilon(k,\beta)\|^2) = \int_{-\infty}^{\infty} \|\varepsilon(k,\beta)\|_2^2 f\{\varepsilon(k,\beta)\} d\varepsilon(k,\beta) \tag{9}$$

The density, 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

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

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. As already mentioned, the AIC will be used later to choose the index.

Since the value of $\mathrm{TM}(K)_t$ 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 $\mathrm{TM}(K)$. 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 cryptos in the crypto universe for which prices are available. The cryptos are weighted by their market capitalization.*

This changes (5) to

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

with k_{max} the maximum number of cryptos with available prices and (8) to

$$\begin{aligned} \min_{k,\beta} \|\widehat{\varepsilon}(k,\beta)\|^2 = \|\varepsilon(k_{max})^{TM} - \varepsilon(k,\beta)^{CRIX}\|^2 \\ \text{s.t.: } 1 \leq k \leq k^u \\ k^u \in [1, k_{max}] \\ s \in [1, k_{max} - k] \\ \beta^{1\times(k+s)} = (1, \ldots, 1, \beta_{k+1}, \ldots, \beta_{k+s})^\top \\ \beta_{k+1}, \ldots, \beta_{k+s} \in (-\infty, \infty). \end{aligned} \tag{10}$$

We introduced several constraints with (10). We will search for an index under the classical approach of Laspeyres, where $\beta = 1$. We include $\beta_{k+1}, \ldots, \beta_{k+s}$ to evaluate if adding s more assets to the index explains the difference between $\varepsilon(k_{max})^{TM}$ and $\varepsilon(k,\beta)^{CRIX}$ better. The first k assets won't be adjusted by a parameter, so no parameter estimation is necessary. This makes the first term a constant. The parameters of the next s assets have to be estimated.

A number of criteria are applicable. In this context: cross validation (CV), full cross validation (FCV), Generalized Cross Validation (GCV), Generalized Full Cross Validation (GFCV), Mallows' C_p, Akaike's Final Prediction Error (FPE), Shibata (SH), AIC, BIC and Hannan Quinn (HQ), Droge (2006). The first one, CV, see Stone (1974), is a widely used criterion

$$\text{CV}(k,\beta) = T^{-1}\sum_{i=1}^{T}\{\varepsilon(k_{max})_t^{TM} - \varepsilon(k,\beta)_{-t}^{CRIX}\}^2 \tag{11}$$

where $\varepsilon(k,\beta)_{-t}^{CRIX}$ is the estimate of $\varepsilon(k,\beta)_t^{CRIX}$ without the observation t. It is however not suitable as FCV in this context because CV does not involve any penalty for the number of constituents, Bunke et al. (1999). The GCV criterion, see Craven and Wahba (1978), is defined as

$$\text{GCV}\{\widehat{\varepsilon}(k,\beta), s\} = \frac{T^{-1}\sum_{t=1}^{T}\widehat{\varepsilon}(k,\beta)_t^2}{(1 - T^{-1}s)^2} \tag{12}$$

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 GCV was shown in several frameworks. The GFCV, see Droge (1996):

$$\text{GFCV}\{\widehat{\varepsilon}(k,\beta), s\} = T^{-1}\sum_{t=1}^{T}\widehat{\varepsilon}(k,\beta)_t^2(1 + T^{-1}s)^2 \tag{13}$$

is an alteration.

A further score, SH,

$$\text{SH}\{\widehat{\varepsilon}(k,\beta), s\} = \frac{T + 2s}{T^2}\sum_{t=1}^{T}\widehat{\varepsilon}(k,\beta)_t^2, \tag{14}$$

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

Mallows (1973)' C_p:

$$C_p\{\widehat{\varepsilon}(k,\beta),s\} = \frac{\sum_{t=1}^{T}\widehat{\varepsilon}(k,\beta)_t^2}{\widehat{\sigma}(k,\beta)^2} - T + 2\cdot s \tag{15}$$

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

$$\mathrm{FPE}\{\widehat{\varepsilon}(k,\beta),s\} = \frac{T+s}{(T-s)T}\sum_{t=1}^{T}\widehat{\varepsilon}(k,\beta)_t^2, \tag{16}$$

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 further criteria use more information by depending on the maximum likelihood, derived by

$$L\{\widehat{\varepsilon}(k,\beta)\} = \max_{\beta}\prod_{t} f\{\widehat{\varepsilon}(k,\beta)_t\}, \tag{17}$$

where f, in (9), represents the density of the $\widehat{\varepsilon}(k,\beta)_t$ over all t. The first one is the AIC which is defined to be

$$\mathrm{AIC}\{\widehat{\varepsilon}(k,\beta),s\} = -2\log L\{\widehat{\varepsilon}(k,\beta)\} + s\cdot 2, \tag{18}$$

Akaike (1998). If the true model is of finite dimension, then neither FPE nor AIC are consistent, compare Hurvich and Tsai (1989). But Shibata (1983) showed the asymptotic efficiency of Mallows' C_p, FPE and AIC under the assumption of an infinite number of regression variables or an increasing number of regression variables with the sample size.

On the other hand, the BIC, defined as

$$\mathrm{BIC}\{\widehat{\varepsilon}(k,\beta),s\} = -2\log L\{\widehat{\varepsilon}(k,\beta)\} + s\cdot\log(T), \tag{19}$$

see Schwarz (1978), is consistent in choosing the true model, Nishii (1984). A further consistent criterion is the one proposed by Hannan and Quinn (1979), defined as

$$\mathrm{HQ}\{\widehat{\varepsilon}(k,\beta),s\} = -2\log L\{\widehat{\varepsilon}(k,\beta)\} + 2s\cdot\log\{\log(T)\}. \tag{20}$$

We'll evaluate now which criteria to use for our purpose. Since CRIX is to be a benchmark model, all possible models under certain restrictions for the number of parameters are included in the test set, $\Theta_{AIC} = \{\mathrm{CRIX}(k_1,\beta),\mathrm{CRIX}(k_2,\beta),\dots\}$, where $k_1,k_2,\dots$ are predefined values. Recall that the intention behind CRIX is to discover the best model to describe the data (benchmark) under a squared loss function.

Define the loss function in (9) for $\widehat{\varepsilon}(k,\beta)$,

$$R_T\{\widehat{\varepsilon}(k,\beta)\} = \mathsf{E}(||\widehat{\varepsilon}(k,\beta)||^2), \tag{21}$$

and define the number of constituents which minimize the risk in $R_T(k,\beta)$ as k^* and s^* for the model set Θ, Shibata (1983). For this paragraph, consider $\widehat{\varepsilon}(k,\beta)\sim N(0,\hat{\sigma}(k,\beta)^2)$. k^* and s^* will be interpreted as the number of constituents which balance the bias and

variance, define

$$H_T\{\widehat{\varepsilon}(k,\beta),s\} = ||\widehat{\varepsilon}(k,\beta)||^2 + s\hat{\sigma}(k,\beta)^2. \tag{22}$$

Mean efficiency shall be defined as

$$\text{eff}(\Theta) = H_T(\widehat{\varepsilon}(k^*,\beta),s^*)/R_T(\Theta). \tag{23}$$

A criteria is defined to be asymptotic mean efficient if

$$\text{a.eff}(\Theta) = \liminf_{T\to\infty} H_T\{\widehat{\varepsilon}(k^*,\beta),s^*\}/R_T(\Theta) = 1 \tag{24}$$

This result holds if the number of constituents , k^* and s^*, increases with T, Shibata (1983). Of course, this result was derived under the assumption of normally distributed errors. Since we are estimating the distribution non parametrically, this result might not hold. For example, Boisbunon et al. (2013) investigate that the result for gaussian distributed errors should hold for spherically symmetric and elliptically contoured distributions too. This leads us to the conclusion that asymptotic optimality might be still given in this case. We oracle so, because for infinitely many observations, the nonparametric estimator tends to the true distribution, Härdle et al. (2004).

The assumption of k^* and s^* increasing with T is plausible in this case since longer time horizons T would include cryptos which aren't part of shorter ones due to bankruptcy or since they haven't been found yet. Both lead to more complexity. It follows that all of the asymptotically optimal criteria would lead to a mean efficient model choice in terms of squared risk for a given selection of models which fits the intention to discover a best model. It remains to find the suitable one.

Define the characteristic function as

$$\varphi(t) = \int_{-\infty}^{\infty} \exp(\mathbf{i}tx) f(x) dx \tag{25}$$

with $\mathbf{i} \in \mathbb{C}$ and $t \in \mathbb{R}$. The Fourier inversion theorem states (Shephard (1991)):

Theorem 1. *Suppose g and φ are integrable in the Lebesgue sense and*

$$\varphi(t) = \int_{-\infty}^{\infty} \exp(\mathbf{i}tx) g(x) dx, \tag{26}$$

then

$$g(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp(-\mathbf{i}tx) \varphi(t) dt. \tag{27}$$

holds everywhere.

The moment generating function is defined as

$$M(t) = \int_{-\infty}^{\infty} \exp(tx) f(x) dx. \tag{28}$$

If the moments generating function exists, it holds

$$\varphi(t) = M(\mathbf{i}t). \tag{29}$$

We see that the characteristic function depends on the moment generating function of $\widehat{\varepsilon}$. Most of the asymptotically optimal criteria depend on the empirical versions of the first

two moments of $\widehat{\varepsilon}$. Just the AIC uses the full distribution via the likelihood and therefore all the moments. This makes its information basis richer. For the derivation of the number of index members of CRIX, we will use the AIC, because it uses the most information compared to the other asymptotically optimal criteria: it is the only one which depends on the likelihood.

To decide with 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,

$$\|\widehat{\varepsilon}(k_j, \beta)\|^2 = \|\varepsilon(k_{max})^{TM} - \varepsilon(k_j, \beta)^{CRIX}\|^2, \tag{30}$$

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, CRIX(k_j, β), have different numbers of constituents which fulfill $k_1 < k_2 < k_3 < \cdots$, where $k_j = k_1 + s\dot{(}j - 1)$. Therefore, the number of constituents between the indices are equally spaced. By definition both information criteria evaluate the differences, $\widehat{\varepsilon}(k_j, \beta)$, between the candidates and the TMI with the respective likelihood $L\{\widehat{\varepsilon}(k_j, \beta), s\}$, see Equations (18) and (19). This procedure implies that the IC 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. We expect assets with a higher market capitalization to have a higher influence on the AIC, so we formulated the following theorem:

Theorem 2. *The rate of improvement of the AIC depends on the relative value of an asset in the market.*

The proof for the Theorem 2 is given in the Appendix, 10.1, 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, see e.g. Deutsche Boerse AG (2013) or MEXBOL (2013),

$$\arg\max_i \sum_{j=1}^{k} P_{j,i,t_l^-} Q_{j,i,t_l^-}, \quad i \in \{1, \ldots, K\}. \tag{31}$$

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

For the sorting of the index constituents by highest market capitalization, we just rely on the closing data of the last day of a month. 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.

Since the differences between the TMI(k_{max}) and CRIX(k_j, β) are caused over time by the missing time series in CRIX(k_j, β), the independence assumption of the $\widehat{\varepsilon}(k_j, \beta)$ for all j can not be fulfilled by construction. But Györfi et al. (1989) give arguments that under certain conditions, the rate of convergence is essentially the same as for an independent sample. Since the same data are used to estimate f^j and the information criterion, a "leave-one-out" cross-validation procedure is performed in order to have in-sample data for the calculation of the density and pseudo-out-of-sample data for the information criterion, hence new observations; see Boisbunon et al. (2013). Summarizing the described procedure, results to:

1. At time point $T + 1$, construct TMI(k_{max})
2. Set $j = 1$

3. Construct CRIX$(k_j, 1)$ and CRIX(k_j, β), $j = 1, 2, 3, \ldots,\ k_1 < k_2 < k_3 < \cdots$
4. Compute $\widehat{\varepsilon}(k_j, \beta)$ and $\widehat{\varepsilon}(k_j, 1)$
5. Kernel density estimation for density $f^j(\widehat{\varepsilon}(k_j, \beta))$ and $f^j(\widehat{\varepsilon}(k_j, 1))$ with leave-one-out cross validation
   a) Perform for $\widehat{\varepsilon}(k_j, \beta)$ and $\widehat{\varepsilon}(k_j, 1)$, where $\widehat{\varepsilon}$ belongs to the respective return series:
   b) Construct T datasets $\widehat{\varepsilon}_{-t} = \{\ldots, \widehat{\varepsilon}_{t-1}, \widehat{\varepsilon}_{t+1}, \ldots\}$, leaving out $\widehat{\varepsilon}_t$.
   c) Compute the Kernel Density Estimator (KDE) for each $\widehat{\varepsilon}_{-t}$.
   d) Compute the log likelihood (18) for $\widehat{\varepsilon}_t$ with KDE for $\widehat{\varepsilon}_{-t}$.
   e) Sum the log likelihoods
6. Derive AIC$\{\widehat{\varepsilon}(k_j, \beta), s\}$ and AIC$\{\widehat{\varepsilon}(k_j, 1), 0\}$
7. If $j = (k_{max} - s)$: 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, see e.g. Deutsche Boerse AG (2013), MEXBOL (2013) and S&P (2014). 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. We choose a monthly reallocation to make sure that CRIX catches the momentum of the cryptocurrency market well. Therefore, on the last day of every month, the cryptos 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.

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 may alter the data in the wrong direction. By LOCF, we imply no change and just do not exclude the crypto. 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 crypto at this time point. CRIX should mimic market changes, therefore an imputation or regression method for the missing data would distort the view of the market.

Before we continue, we summarize the rules which were described so far:

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

   $$\text{AIC}\{\widehat{\varepsilon}(k_j, \beta), 5\} < \text{AIC}\{\widehat{\varepsilon}(k_j, 1), 0\}. \tag{32}$$

   Therefore merely a local optimum will be achieved in most of the cases for $\Theta = \Theta_{AIC}$, in (23). 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:
   The second constructed index is called Exact CRIX (ECRIX). 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), $k = 1, 2, 3, \ldots$ and stops when

   $$\text{AIC}\{\widehat{\varepsilon}(k_j, \beta), 1\} < \text{AIC}\{\widehat{\varepsilon}(k_j, 1), 0\}. \tag{33}$$

3. EFCRIX:
   Since the decision procedures for CRIX and ECRIX terminate when the AIC rises for the first time, Exact Full CRIX will be constructed to visualize whether the decision procedure works fine for the two covered indices. The intention is to have an index which may approach the TMI but only in case even small assets help improve

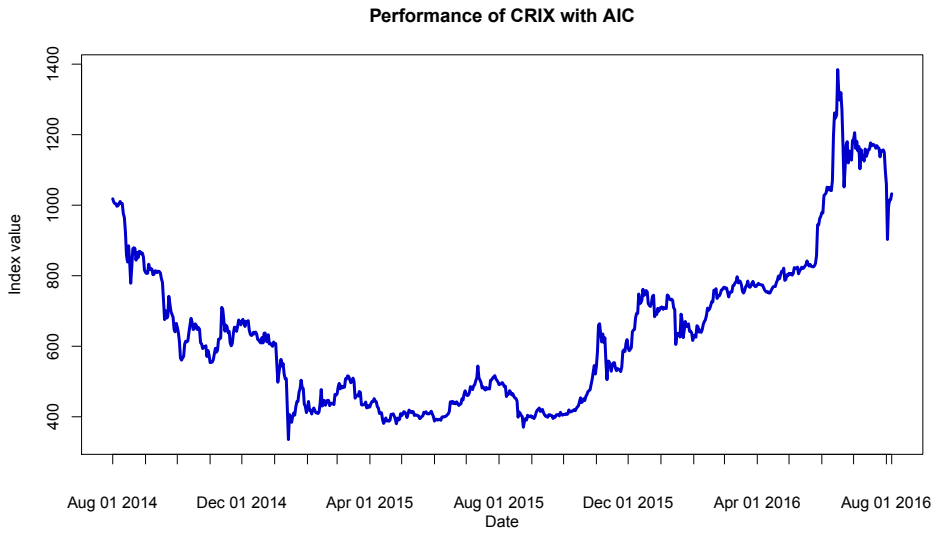

Figure 2: Performance of CRIX

CRIXindex CRIXcode

the view of the total market, a benchmark for the benchmarks. It'll be derived with the AIC procedure, compare Section 3. The decision rule is based on

$$\min_{k_j,\beta} \text{AIC}\{\widehat{\varepsilon}(k_j,\beta),1\} \tag{34}$$

for $\Theta = \Theta_{AIC}$, in (23). This index computes the AIC for every possible number of constituents and what is chosen is the number where the AIC becomes minimal.

6 Performance analysis

The indices CRIX, ECRIX, EFCRIX have been proposed to give insight into the crypto market. Our RDC crypto database covers data for 331 cryptocurrencies, 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 crypto in the time period from 2014-04-01 to 2016-08-06. Crypto exchanges are open on the weekends, therefore data for weekend closing prices exist. Since crypto 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 Chapter 4 will come into play.

Figure 2 shows the performance of CRIX, and Figure 3 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 3. For comparison, we provide the number of constituents under the other discussed model selection criteria 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, 3. Apparently the methodology of EFCRIX

causes its number of constituents to become close to maximal in every 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 GCV, GFCV and SH tend to choose more or the same number of constituents than AIC. Also all three criteria suggest the same result. FPE results in much more constituents than any other criteria for CRIX and ECRIX. C_p stops at the initial value for CRIX and ECRIX, just for EFCRIX it differs. For CRIX, ECRIX and EFCRIX, AIC chooses less constituents compared to all other criteria, except C_p which terminates very early.

	AIC	GCV	GFCV	SH	Cp	FPE
CRIX	18.4500	3.4049	3.4049	3.4049	25.7485	0.0312
ECRIX	444.2490	386.8260	386.8260	386.8260	768.4531	0.0324
EFCRIX	0.0314	0.0324	0.0324	0.0324	165.9407	0.0324

Table 1: Comparison of CRIX, ECRIX, EFCRIX, derived under different penalizations, against TMI under Mean Squared Error

	AIC	GCV	GFCV	SH	Cp	FPE
CRIX	0.9891	0.9946	0.9946	0.9946	0.9864	1.0000
ECRIX	0.9511	0.9592	0.9592	0.9592	0.9117	1.0000
EFCRIX	1.0000	1.0000	1.0000	1.0000	0.9728	1.0000

Table 2: Comparison of CRIX, ECRIX, EFCRIX, derived under different penalizations, against TMI under Mean Directional Accuracy

Since the indices CRIX and ECRIX are just optimized until a local optimum, they are expected to perform less optimal than the EFCRIX against the TMI. Table 1 gives the Mean Square Error (MSE) and Table 2 the Mean Directional Accuracy (MDA), defined as

$$\mathrm{MSE}\{\mathrm{CRIX}(k)\} = \frac{1}{T}\sum_{t=1}^{T}\{\mathrm{CRIX}(k)_t - \mathrm{TMI}(k_{\max})_t\}^2 \tag{35}$$

$$\begin{aligned}\mathrm{MDA}\{\mathrm{CRIX}(k)\} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{I}[&\mathrm{sign}\{\mathrm{TMI}(k_{\max})_t - \mathrm{TMI}(k_{\max})_{t-1}\} \\ &= \mathrm{sign}\{\mathrm{CRIX}(k)_t - \mathrm{CRIX}(k)_{t-1}\}]\end{aligned} \tag{36}$$

where $\mathbf{I}(\cdot)$ is the indicator function and $\mathrm{sign}(\cdot)$ gives the sign of the respective equation. The recalibration of the indices on the recalculation date is important for the computation of the MSE, since altering the constituents may change the future development in terms of MSE. The MDA is insensitive to the recalibration. Apparently EFCRIX 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 EFCRIX. Additionally, ECRIX performs worse than CRIX in terms of MSE and MDA. Comparing all the model selection criteria, we see that FPE has the best performance in terms of MSE and MDA. This is not surprising since this criteria selected the most constituents. The more important risk criteria is the MDA, since the direction of the TMI shall be adequately modeled by the criteria. Table 2 shows that the AIC either performs second

worst or performs equally good to the other criteria. But the MDA is close to the one of the outperforming ones. The reason for the lower value is the lower number of constituents. Interestingly, AIC outperforms in terms of MSE and MDA for EFCRIX, which choses the global optimum.

CRIX was constructed with steps of five which is common in practice, but this analysis showed that ECRIX would work well for the crypto market too. Additionally, the analysis showed that it is indeed unnecessary to choose the global optimal AIC. Even a local optimum and a much lower number of constituents is able to mimic the market movements very well in terms of the MDA. 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, doesn't lead the market. Other cryptocurrencies are important for the market movements too.

We are aware that the AIC didn't outperform the other criteria in terms of MDA and MSE for the local optimum. But for the global optimum, AIC outperforms while having sometimes less constituents. We conclude that the AIC works best in search of the global optimum, but for the local optimum we observed a diverging picture. But the AIC still performed very well. Due to the theoretical choice for the AIC and its global optimal behavior, we continue working with the AIC.

	CRIX						ECRIX						EFCRIX					
	AIC	GCV	GFCV	SH	Cp	FPE	AIC	GCV	GFCV	SH	Cp	FPE	AIC	GCV	GFCV	SH	Cp	FPE
1	10	10	10	10	5	40	1	4	4	4	1	41	41	41	41	41	1	41
2	30	30	30	30	5	115	5	14	14	14	1	117	98	117	117	117	4	117
3	30	30	30	30	5	160	4	4	4	4	1	162	161	162	162	162	1	162
4	5	45	45	45	5	190	3	3	3	3	1	192	187	192	192	192	14	192
5	15	15	15	15	5	205	1	6	6	6	1	208	206	208	208	208	1	208
6	20	35	35	35	5	215	6	16	16	16	1	215	215	215	215	215	20	215
7	5	5	5	5	5	220	1	1	1	1	1	221	220	221	221	221	10	221
8	15	65	65	65	5	170	3	3	3	3	1	167	153	167	167	167	5	167
9	5	5	5	5	5	220	2	5	5	5	1	223	219	223	223	223	3	223

Table 3: Comparison of AIC, GCV, GFCV, SH, Cp and the FPE method for the selection of the number of index constituents for the CRIX, ECRIX and EFCRIX in the 9 periods

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. We tested the proposed methodology 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. Following Definition 4, we define all available stocks as the TMI and apply our new method 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

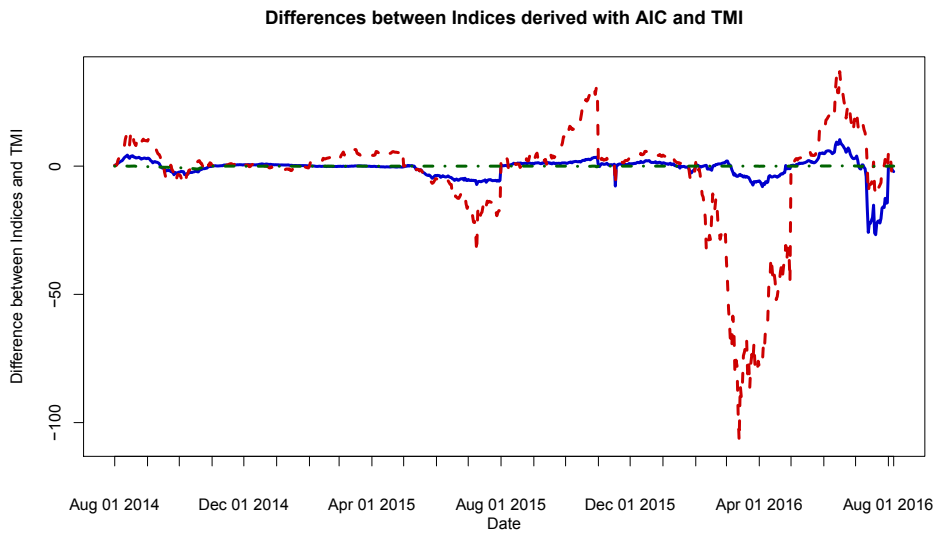

Figure 3: Realized difference between TMI and CRIX (solid), ECRIX (dashed), EFCRIX (dotdashed)

CRIXfamdiff CRIXcode

reallocation dates, the 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 20000616 until 20151218. We took all stocks which are German companies and are traded on XETRA. 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.

Figure 4 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 Section 10. Figure 4 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, 2006 and 2011.

The computation of the MSE and MDA, see Table 4, 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.

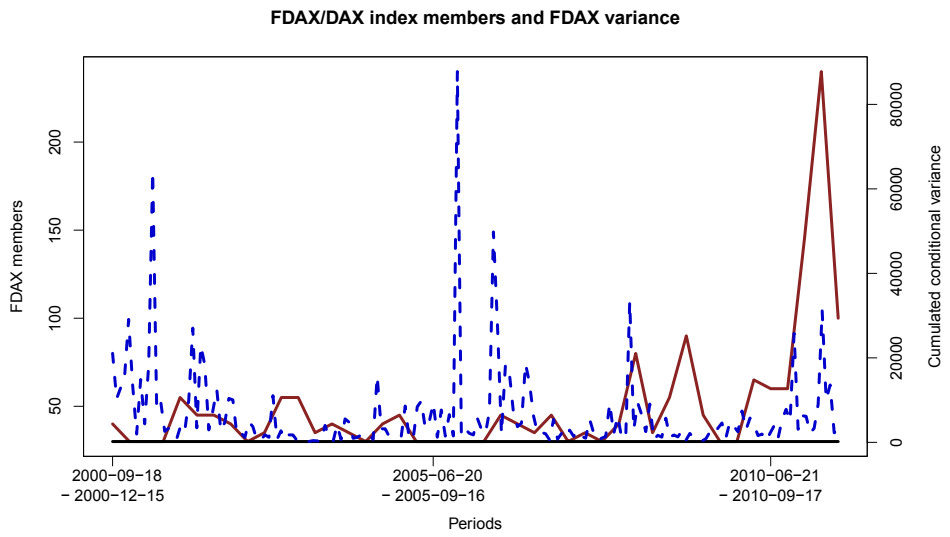

Figure 4: Number of constituents of FDAX (solid), DAX (horizontal solid) and cumulated monthly variance of FDAX (dashed)

CRIXdaxmembersvar CRIXcode

	MSE	MDA
FDAX vs. TMI	189.47	0.96
DAX vs. TMI	73769.07	0.92

Table 4: Comparison of DAX with CRIX methodology (FDAX) and rescaled DAX against TMI

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 crypto market Bitcoin is such a dominant asset. The CRIX methodology could help to circumvent arbitrary rules and develop an index to represent the market accurately.

The data were fetched from Datastream for the period 19960601 until 20150529 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, we recalculated the index 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 5. 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. Figure (5) shows the number of index members of the FIPC compared to the IPC. Obviously, the methodology also suggests using more than 35 index members

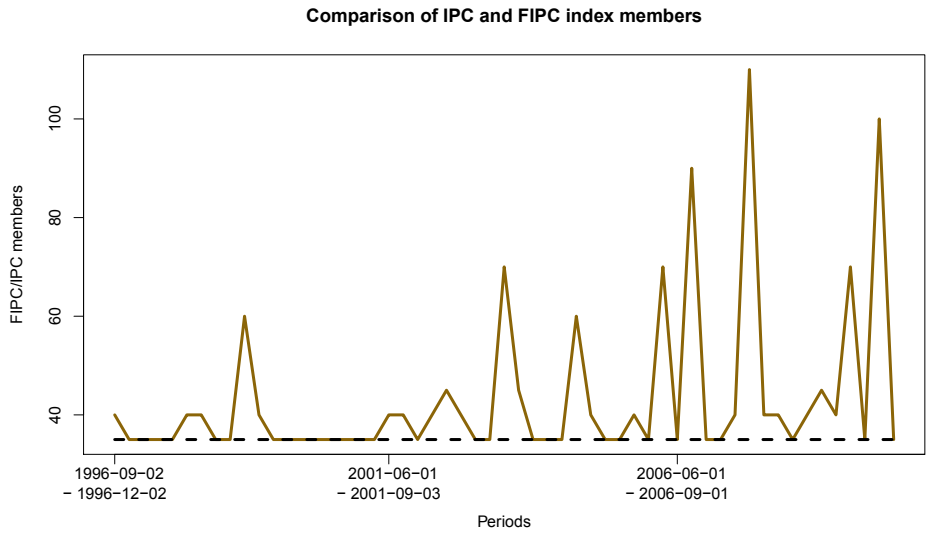

Figure 5: Number of constituents of FIPC (solid) and IPC (dashed) in the respective periods

CRIXipcmembers CRIXcode

most of the time which is the number of members of the IPC.

	MSE	MDA
FIPC vs. TMI	319.25	0.97
IPC vs. TMI	27177417.58	0.91

Table 5: Comparison of IPC with CRIX methodology (FIPC) and rescaled IPC against TMI

9 Conclusion

The movements of cryptocurrencies are very different from each other, Elendner et al. (2016). So studying the entire market of cryptocurrencies requires an instrument which adequately captures and displays the market movements, an index. But index construction for cryptocurrencies requires a new methodology to find the right number of index members. Innovative markets, like the one for cryptocurrency's, 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 on `hu.berlin/crix`. 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 crypto market - momentarily dominated by Bitcoin - still needs a representative index since Bitcoin does not lead the market. 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.

References

Akaike, H. (1998). "Information Theory and an Extension of the Maximum Likelihood Principle". *Selected Papers of Hirotugu Akaike*. Ed. by E. Parzen, K. Tanabe, and G. Kitagawa. Springer Series in Statistics. Springer New York, pp. 199–213.

Akaike, H. (1970). "Statistical predictor identification". *Annals of the Institute of Statistical Mathematics* 22.1, pp. 203–217.

Arlot, S. and A. Celisse (2010). "A survey of cross-validation procedures for model selection". *Statistics Surveys* 4, pp. 40–79.

Boisbunon, A., S. Canu, D. Fourdrinier, W. Strawderman, and M. T. Wells (2013). "AIC, Cp and estimators of loss for elliptically symmetric distributions". *arXiv:1308.2766 [math, stat]*.

Bollerslev, T. (1986). "Generalized autoregressive conditional heteroskedasticity". *Journal of Econometrics* 31.3, pp. 307–327.

Bunke, O., B. Droge, and J. Polzehl (1999). "Model Selection, Transformations and Variance Estimation in Nonlinear Regression". *Statistics* 33, pp. 197–240.

Chen, S., C. Y.-H. Chen, W. K. Härdle, T. M. Lee, and B. Ong (2016). "A first econometric analysis of the CRIX family". *SFB 649 Discussion Paper forthcoming in Digital Banking and Internet Finance*.

Craven, P. and G. Wahba (1978). "Smoothing noisy data with spline functions". *Numerische Mathematik* 31.4, pp. 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. and L. Györfi (1985). *Nonparametric Density Estimation The L1 View*. Wiley.

Droge, B. (1996). "Some Comments on Cross-Validation". *Statistical Theory and Computational Aspects of Smoothing*. Ed. by W. K. Härdle and M. G. Schimek. Contributions to Statistics. Physica-Verlag HD, pp. 178–199.

Droge, B. (2006). "Asymptotic properties of model selection procedures in linear regression". *Statistics* 40.1, pp. 1–38.

EconoTimes (2016). "Japans Cabinet Approves New Bitcoin Regulations". *econotimes.com*.

Elendner, H., S. Trimborn, B. Ong, and T. M. Lee (2016). "The Cross-Section of Crypto-Currencies as Financial Assets: An Overview". *SFB 649 Discussion Paper forthcoming in Digital Banking and Internet Finance*.

Epanechnikov, V. (1969). "Non-Parametric Estimation of a Multivariate Probability Density". *Theory of Probability & Its Applications* 14.1, pp. 153–158.

FTSE (2016). “FTSE UK Index Series”. *www.ftse.com*.

Györfi, L., W. K. Härdle, P. Sarda, and P. Vieu (1989). *Nonparametric Curve Estimation from Time Series*. Ed. by L. Györfi, W. K. Härdle, P. Sarda, and P. Vieu. Lecture Notes in Statistics 60. Springer New York.

Hall, P. (1987). “On Kullback-Leibler Loss and Density Estimation”. *The Annals of Statistics* 15.4, pp. 1491–1519.

Hannan, E. J. and B. G. Quinn (1979). “The Determination of the Order of an Autoregression”. *Journal of the Royal Statistical Society. Series B (Methodological)* 41.2, pp. 190–195.

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

Härdle, W. K. and S. Trimborn (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*. 3. Edition. London: Institute of Economic Affairs.

Horton, N. J. and K. P. Kleinman (2007). “Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models”. *The American Statistician* 61.1, pp. 79–90.

Hurvich, C. M. and C.-L. Tsai (1989). “Regression and time series model selection in small samples”. *Biometrika* 76.2, pp. 297–307.

Janßen, B. and B. Rudolph (1992). “Der Deutsche Aktienindex DAX”. *Fritz Knapp Verlag*.

Kanazawa, Y. (1993). “Hellinger distance and Kullback—Leibler loss for the kernel density estimator”. *Statistics & Probability Letters* 18.4, pp. 315–321.

Kawa, L. (2015). “Bitcoin Is Officially a Commodity, According to U.S. Regulator”. *Bloomberg.com*.

Kristoufek, L. (2014). “What are the main drivers of the Bitcoin price? Evidence from wavelet coherence analysis”. *arXiv:1406.0268 [physics, q-fin]*.

Mallick, H. and N. Yi (2013). “Bayesian Methods for High Dimensional Linear Models”. *Journal of Biometrics & Biostatistics* 1.

Mallows, C. L. (1973). “Some Comments on Cp”. *Technometrics* 15.4, pp. 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”. *The Annals of Statistics* 12.2, pp. 758–765.

NYSE (2015). "StrataQuant Index Family". *www.nyse.com*.

Potapov, A., J. R. Muirhead, S. R. Lele, and M. A. Lewis (2011). "Stochastic gravity models for modeling lake invasions". *Ecological Modelling* 222.4, pp. 964–972.

Reid, F. and M. Harrigan (2013). "An Analysis of Anonymity in the Bitcoin System". *Security and Privacy in Social Networks*. Ed. by Y. Altshuler, Y. Elovici, A. B. Cremers, N. Aharony, and A. Pentland. Springer New York, pp. 197–223.

Ron, D. and A. Shamir (2013). "Quantitative Analysis of the Full Bitcoin Transaction Graph". *Financial Cryptography and Data Security*. Ed. by A.-R. Sadeghi. Lecture Notes in Computer Science 7859. Springer Berlin Heidelberg, pp. 6–24.

Schwarz, G. (1978). "Estimating the Dimension of a Model". *The Annals of Statistics* 6.2, pp. 461–464.

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

Shephard, N. G. (1991). "From Characteristic Function to Distribution Function: A Simple Framework for the Theory". *Econometric Theory* 7.4, pp. 519–529.

Shibata, R. (1981). "An Optimal Selection of Regression Variables". *Biometrika* 68.1, pp. 45–54.

Shibata, R. (1983). "Asymptotic mean efficiency of a selection of regression variables". *Annals of the Institute of Statistical Mathematics* 35.1, pp. 415–423.

S&P (2014). "Index Mathematics - Methodology". *us.spindices.com*.

S&P (2015). "Dow Jones Total Stock Market Indices Methodology". *us.spindices.com*.

Stone, M. (1974). "Cross-validatory choice and assessment of statistical predictions". *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 111–147.

Tschorsch, F. and B. Scheuermann (2015). "Bitcoin and Beyond: A Technical Survey on Decentralized Digital Currencies". *IEEE Communications Surveys Tutorials*.

Wand, M. P. and M. C. Jones (1994). "Multivariate plug-in bandwidth selection". *Computational Statistics* 9.2, pp. 97–116.

Wilshire Associates (2015). "Wilshire 5000 Total Market Index Methodology". *wilshire.com*.

Woodroofe, M. (1982). "On Model Selection and the ARC Sine Laws". *The Annals of Statistics* 10.4, pp. 1182–1194.

10 Appendix

10.1 Proof of Theorem 2

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

$$\log L\{\varepsilon(k,\beta)\} = -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log\sigma(k,\beta)^2 - \frac{1}{2\sigma(k,\beta)^2}\sum_{t=1}^{T}\varepsilon(k,\beta)_t^2. \tag{37}$$

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

$$\log L\{\widehat{\varepsilon}(k,\beta)\} = -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log T^{-1}RSS\{\widehat{\varepsilon}(k,\beta)\} - \frac{1}{2T^{-1}RSS\{\widehat{\varepsilon}(k,\beta)\}}RSS\{\widehat{\varepsilon}(k,\beta)\} \tag{38}$$

$$= -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log T^{-1}RSS\{\widehat{\varepsilon}(k,\beta)\} - \frac{T}{2} \tag{39}$$

$$= -\frac{T}{2}\log T^{-1}RSS\{\widehat{\varepsilon}(k,\beta)\} + C \tag{40}$$

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.

$$\mathrm{AIC}\{\widehat{\varepsilon}(k,\beta), s\} = T\log T^{-1}RSS\{\widehat{\varepsilon}(k,\beta)\} + 2\cdot s \tag{41}$$

$$= T\log\hat{\sigma}(k,\beta)^2 + 2\cdot s \tag{42}$$

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

$$\begin{aligned} \mathsf{E}\{\varepsilon(k_{max})^{TM}\} = \mathsf{E}\{\varepsilon(k,\beta)^{CRIX}\} &= 0 \\ t &\in \{1,\ldots,T\} \\ t_l^- &= 0 \\ s &= 1 \end{aligned}$$

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

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

$$
\begin{aligned}
&= \sum_{t=1}^{T}\Bigg[\log\left\{\sum_{i=1}^{k_{max}} P_{it}Q_{i,0}\sum_{i=1}^{k} P_{i,t-1}Q_{i,0}\right\} + \log\left\{1 + \frac{\sum_{i=1}^{k_{max}} P_{it}Q_{i,0}\beta_1 P_{k+1,t-1}Q_{k+1,0}}{\sum_{i=1}^{k_{max}} P_{it}Q_{i,0}\sum_{i=1}^{k} P_{i,t-1}Q_{i,0}}\right\} \\
&\qquad - \log\left\{\sum_{i=1}^{k_{max}} P_{i,t-1}Q_{i,0}\sum_{i=1}^{k} P_{i,t}Q_{i,0}\right\} + \log\left\{1 + \frac{\sum_{i=1}^{k_{max}} P_{i,t-1}Q_{i,0}\beta_1 P_{k+1,t}Q_{k+1,0}}{\sum_{i=1}^{k_{max}} P_{i,t-1}Q_{i,0}\sum_{i=1}^{k} P_{i,t}Q_{i,0}}\right\}\Bigg]^2 \\
&= \sum_{t=1}^{T}\Bigg(\log\left\{\sum_{i=1}^{k_{max}} P_{it}Q_{i,0}\sum_{i=1}^{k} P_{i,t-1}Q_{i,0}\right\} - \log\left\{\sum_{i=1}^{k_{max}} P_{i,t-1}Q_{i,0}\sum_{i=1}^{k} P_{i,t}Q_{i,0}\right\} \\
&\qquad + \left[\log\left\{1 + \frac{\beta_1 P_{k+1,t-1}Q_{k+1,0}}{\sum_{i=1}^{k} P_{i,t-1}Q_{i,0}}\right\} - \log\left\{1 + \frac{\beta_1 P_{k+1,t}Q_{k+1,0}}{\sum_{i=1}^{k} P_{i,t}Q_{i,0}}\right\}\right]\Bigg)^2 \qquad (43)
\end{aligned}
$$

Solving the derivation and writing the terms which do not depend on β_1 as A_t and the last part of (43) as B_t:

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

Since normally distributed error terms are assumed, note that $\beta_1 = \frac{Cov\{\widehat{\varepsilon}(k,1),\varepsilon_{k+1}\}}{Var\{\varepsilon_{k+1}\}}$, where ε_{k+1} is the log return of $P_{i,t}Q_{i,0}$. We see that the change in the variance will depend on the additional variance which the new constituent can explain, see β_1. Furthermore, it depends on the value of $P_{k+1,t}Q_{k+1,0}$ relative to $\sum_{i=1}^{k} P_{i,t}Q_{i,0}$, (43), 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.

SFB 649 Discussion Paper Series 2016

For a complete list of Discussion Papers published by the SFB 649, please visit http://sfb649.wiwi.hu-berlin.de.

001 "Downside risk and stock returns: An empirical analysis of the long-run and short-run dynamics from the G-7 Countries" by Cathy Yi-Hsuan Chen, Thomas C. Chiang and Wolfgang Karl Härdle, January 2016.
002 "Uncertainty and Employment Dynamics in the Euro Area and the US" by Aleksei Netsunajev and Katharina Glass, January 2016.
003 "College Admissions with Entrance Exams: Centralized versus Decentralized" by Isa E. Hafalir, Rustamdjan Hakimov, Dorothea Kübler and Morimitsu Kurino, January 2016.
004 "Leveraged ETF options implied volatility paradox: a statistical study" by Wolfgang Karl Härdle, Sergey Nasekin and Zhiwu Hong, February 2016.
005 "The German Labor Market Miracle, 2003 -2015: An Assessment" by Michael C. Burda, February 2016.
006 "What Derives the Bond Portfolio Value-at-Risk: Information Roles of Macroeconomic and Financial Stress Factors" by Anthony H. Tu and Cathy Yi-Hsuan Chen, February 2016.
007 "Budget-neutral fiscal rules targeting inflation differentials" by Maren Brede, February 2016.
008 "Measuring the benefit from reducing income inequality in terms of GDP" by Simon Voigts, February 2016.
009 "Solving DSGE Portfolio Choice Models with Asymmetric Countries" by Grzegorz R. Dlugoszek, February 2016.
010 "No Role for the Hartz Reforms? Demand and Supply Factors in the German Labor Market, 1993-2014" by Michael C. Burda and Stefanie Seele, February 2016.
011 "Cognitive Load Increases Risk Aversion" by Holger Gerhardt, Guido P. Biele, Hauke R. Heekeren, and Harald Uhlig, March 2016.
012 "Neighborhood Effects in Wind Farm Performance: An Econometric Approach" by Matthias Ritter, Simone Pieralli and Martin Odening, March 2016.
013 "The importance of time-varying parameters in new Keynesian models with zero lower bound" by Julien Albertini and Hong Lan, March 2016.
014 "Aggregate Employment, Job Polarization and Inequalities: A Transatlantic Perspective" by Julien Albertini and Jean Olivier Hairault, March 2016.
015 "The Anchoring of Inflation Expectations in the Short and in the Long Run" by Dieter Nautz, Aleksei Netsunajev and Till Strohsal, March 2016.
016 "Irrational Exuberance and Herding in Financial Markets" by Christopher Boortz, March 2016.
017 "Calculating Joint Confidence Bands for Impulse Response Functions using Highest Density Regions" by Helmut Lütkepohl, Anna Staszewska-Bystrova and Peter Winker, March 2016.
018 "Factorisable Sparse Tail Event Curves with Expectiles" by Wolfgang K. Härdle, Chen Huang and Shih-Kang Chao, March 2016.
019 "International dynamics of inflation expectations" by Aleksei Netšunajev and Lars Winkelmann, May 2016.
020 "Academic Ranking Scales in Economics: Prediction and Imdputation" by Alona Zharova, Andrija Mihoci and Wolfgang Karl Härdle, May 2016.

SFB 649, Spandauer Straße 1, D-10178 Berlin
http://sfb649.wiwi.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".

SFB 649 Discussion Paper Series 2016

For a complete list of Discussion Papers published by the SFB 649, please visit http://sfb649.wiwi.hu-berlin.de.

021 "CRIX an Index for blockchain based currencies" by Simon Trimborn and Wolfgang Karl Härdle, May 2016.

SFB 649, Spandauer Straße 1, D-10178 Berlin
http://sfb649.wiwi.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".