Hoga, Yannick

Article — Published Version

Quantifying the data-dredging bias in structural break tests

Statistical Papers

Provided in Cooperation with:
Springer Nature

Suggested Citation: Hoga, Yannick (2021) : Quantifying the data-dredging bias in structural break tests, Statistical Papers, ISSN 1613-9798, Springer, Berlin, Heidelberg, Vol. 63, Iss. 1, pp. 143-155, https://doi.org/10.1007/s00362-021-01233-4

This Version is available at:
https://hdl.handle.net/10419/286808

Statistical Papers (2022) 63:143–155
https://doi.org/10.1007/s00362-021-01233-4

REGULAR ARTICLE

Quantifying the data-dredging bias in structural break tests

Yannick Hoga[1]

Received: 3 November 2020 / Revised: 19 March 2021 / Accepted: 30 March 2021 /
Published online: 16 April 2021

Abstract
Structural break tests are often applied as a pre-step to ensure the validity of subsequent statistical analyses. Without any a priori knowledge of the type of breaks to expect, eye-balling the data can indicate changes in some parameter, e.g., the mean. This, however, can distort the result of a structural break test for that parameter, because the data themselves suggested the hypothesis. In this paper, we formalize the eye-balling procedure and theoretically derive the implied size distortion of the structural break test. We also show that eye-balling a stretch of historical data for possible changes in a parameter does not invalidate the subsequent procedure that monitors for structural change in new incoming observations. An empirical application to Bitcoin returns shows that taking into account the data-dredging bias, which is incurred by looking at the data, can lead to different test decisions.

Keywords Data-dredging bias · Hypothesis test · Monitoring · Structural breaks

JEL Classification C12 (Hypothesis Testing) · C18 (Methodological Issues)

1 Motivation

The importance of plotting the data as a first step of a statistical analysis is stressed in numerous textbooks (e.g., Ruppert and Matteson 2015; Brockwell and Davis 2016). For instance, Brockwell and Davis (2016, p. 12) recommend time series plots to check whether there are 'any apparent sharp changes in behavior'. If such a change is present in the data, yet is ignored in the subsequent analysis, the conclusions drawn from the data may be invalid (see, e.g., Baltagi et al. 2013; Demetrescu and Hanck 2013; Har-

The author is grateful to two referees and Christoph Hanck for their valuable comments and suggestions. This work was supported by the German Research Foundation (DFG) under Grant HO 6305/1-1.

✉ Yannick Hoga
yannick.hoga@vwl.uni-due.de

[1] Faculty of Economics and Business Administration, University of Duisburg-Essen, Universitätsstraße 12, 45117 Essen, Germany

vey et al. 2013; Xu 2015). For instance, Xu (2015) shows that if a structural break in the error variances is ignored, standard tests for the the constancy of regression coefficients suffer from size distortions—even asymptotically. To avoid such misleading test results, applying a formal structural break test is typically recommended as a pre-step to the actual statistical analysis of the data.

However, one problem with the recommendation to look for 'any apparent sharp changes in behavior' is that the decision to apply the structural break test has been informed by the data. Hence, any change point test, if it is to be valid, needs to hold size *conditional* on having looked at the data. Of course, such tests are exclusively constructed to hold size *unconditionally* and, hence, suffer from size distortions if applied otherwise. The first main aim of this paper is to quantify these size distortions for structural break tests that are applied *conditional* on large deviations being observed in the data. We show that these size distortions can become so large that a true null is rejected with certainty.

This changes when one moves from a structural break context, where all data are available in advance, to a monitoring context, where—after having observed some training data—the data become available 'as you go' for sequential tests of parameter stability. Of course, ex ante it may be unclear precisely which parameters to monitor for constancy. One possibility may be to monitor a parameter, whose estimates have fluctuated somewhat in the training data. Then, similarly as above, the monitoring procedure needs to hold size *conditionally* on large fluctuations in the training data. The second main aim of this paper is to show that, unlike for one-shot structural break tests, this is indeed the case.

We mention that the issue of investigating the data to 'decide' which hypotheses to test is an old one. Selvin and Stuart (1966) call this 'hunting', because the investigator hunts for hypotheses to be tested based on the data. They illustrate the practice with Pearson's χ^2 goodness-of-fit (GoF) test. In applications, a possible distribution for the data is usually not chosen *ex ante*, but *ex post* by eye-balling a histogram. This practice is to some extent unavoidable, as in our structural break example. Selvin and Stuart (1966) conclude with regard to hunting that 'the only criticism to be made is of the delusion that one has to pay no price for the sport.' Interestingly, while in general the bias incurred by hunting—as in the GoF example—is hard to quantify, it is possible for monitoring procedures and (up to a single unknown parameter) also for structural break tests.

Even outside the statistical literature, testing hypothesis on the same data that inspired it is known to be harmful. In the social sciences, Kerr (1998) calls this HARKing (Hypothesizing After the Results are Known) and defines it as presenting a post hoc hypothesis (i.e., one informed by the collected data) as an a priori hypothesis. Among others, Simmons et al. (2011) and Gelman and Loken (2014) point out that data analyses in the social sciences are often driven by the observed data, and show that this contributes—among other factors—to the prevalence of false negatives in published work (which is a finding that has led to the so-called *replication crisis*). This is as in our one-shot structural break setting, where the data-inspired hypothesis ('there is a structural break in the data') is more likely to be accepted even when wrong, producing a false negative.

The remainder of the paper proceeds as follows. Section 2 states and discusses the main theoretical results of this paper. The proofs of these results are deferred to the Appendix. An empirical application to Bitcoin returns in Sect. 3 demonstrates that if the decision to apply a structural break test is made *conditional* on the data, different results may be obtained when this fact is taken into account. The final Sect. 4 concludes.

2 Main results

2.1 Structural break tests

Let $X_1, \ldots, X_n$ denote the (possibly multivariate) observations to be tested for a structural break in some scalar parameter γ_i of their respective distribution function F_i ($i = 1, \ldots, n$). For instance, γ_i may be the mean or the variance of a component series, or it may denote the correlation or tail dependence coefficient of two components of the series. Structural break tests for these parameters are well-established (see, e.g., Inclan and Tiao 1994; Vogelsang 1998; Wied et al. 2012; Hoga 2018). Interest in change point tests focuses on the null hypothesis

$$\mathcal{H}_0^S \; : \; \gamma = \gamma_1 = \cdots = \gamma_n,$$

i.e., the constancy of the parameter γ over time.

Let $\widehat{\gamma}(0, t)$ denote an estimate of γ based on the subsample $X_1, \ldots, X_{\lfloor nt \rfloor}$. Here, $\lfloor \cdot \rfloor$ rounds down to the nearest integer. Further, let $D[0, 1]$ denote the space of real-valued functions on $[0, 1]$ that possess left-hand limits and are right-continuous (Davidson 1994). We make the following high-level assumption under $\mathcal{H}_0^S$:

Assumption 1 It holds that, as $n \to \infty$,

$$t\sqrt{k_n}\left[\widehat{\gamma}(0, t) - \gamma\right] \xrightarrow{d} \sigma W(t) \qquad in D[0, 1],$$

where $\sigma > 0$, $k_n \to \infty$, and $\{W(t)\}_{t\in[0,1]}$ denotes a standard Brownian motion.

Such a functional central limit theorem has been shown to hold for many parameters. For the leading case $k_n = n$, it holds (e.g.) for the mean (Davidson 1994), the variance (Wied et al. 2012), correlation (Wied et al. 2012) and Kendall's tau (Dehling et al. 2017). For extreme value quantities, whose estimators typically depend only on a vanishing fraction of the sample, a scaling different from $\sqrt{n}$ is required in Assumption 1. For instance, Assumption 1 holds for the tail index (Hoga 2017a), an extreme quantile estimator (Hoga 2017b) and the tail dependence coefficient (Hoga 2018) for some $k_n = o(n)$. For feasible break testing, the nuisance parameter σ in Assumption 1 needs to be estimated consistently. To that end, we impose the following assumption under $\mathcal{H}_0^S$:

Assumption 2 There exists an estimator $\widehat{\sigma}$ satisfying $\widehat{\sigma} \xrightarrow{p} \sigma$, as $n \to \infty$.

The usual recursive test statistic for testing $\mathcal{H}_0^S$ is

$$T_n = \frac{1}{\widehat{\sigma}} \sup_{t\in[0,1]} \left| t\sqrt{k_n}\left[\widehat{\gamma}(0,t) - \widehat{\gamma}(0,1)\right]\right| \xrightarrow{d} \sup_{t\in[0,1]} |W(t) - tW(1)|, \qquad n \to \infty, \tag{1}$$

where the convergence follows from Assumptions 1 and 2 and the continuous mapping theorem together with Slutsky's theorem (Davidson 1994). Here, fluctuations in the recursive parameter estimates $\widehat{\gamma}(0,t)$ that deviate 'too much' from the full-sample estimate $\widehat{\gamma}(0,1)$ are taken as evidence against the null. Based on the $(1-\alpha)$-quantile c_α^S of the limiting distribution in (1), we reject $\mathcal{H}_0^S$ at significance level $\alpha \in (0,1)$ if $T_n > c_\alpha^S$, since from (1)

$$\Pr\left\{T_n > c_\alpha^S\right\} \longrightarrow \alpha, \qquad n \to \infty. \tag{2}$$

However, often the decision to apply a structural break test (usually to validate the intended subsequent statistical analysis) is not made before the data have been collected. Rather, as pointed out in the Motivation, it is made afterwards based on having observed some large deviations in the series. This eye-balling may be formalized as testing if and only if $t|\widehat{\gamma}(0,t) - \widehat{\gamma}(0,1)|/\widehat{\sigma} > \delta/\sqrt{k_n}$ for some $t \in [0,1]$, where the 'if and only if'-part of course constitutes a crude approximation. Here, the prefactor t discounts a large deviation that is based on very few data points for small t; the inclusion of $\sqrt{k_n}$ reflects smaller expected fluctuations in larger samples; $\widehat{\sigma}^2$ estimates the asymptotic variance of $\widehat{\gamma}(0,1)$ and, hence, reflects estimation uncertainty; finally, $\delta > 0$ determines the (unknown) sensitivity of the visual inspection. In other words, δ is the parameter for which eye-balling for changes can be best approximated by the conditioning event $\{\sup_{t\in[0,1]} t|\widehat{\gamma}(0,t) - \widehat{\gamma}(0,1)|/\widehat{\sigma} > \delta/\sqrt{k_n}\}$.

Of course, this approximation of the eye-balling heuristic may not be perfect, but we argue that it is close enough to be interesting. For instance, plots of $t \mapsto \widehat{\gamma}(0,t)$ are often used as a diagnostic tool indicating structural change; see, e.g., Quintos et al. (2001, Fig. 3) or Wied et al. (2012, Fig. 1). For obvious reasons, we call the above procedure 'formalized' eye-balling.

Now, carrying out the structural break test if only if $t|\widehat{\gamma}(0,t) - \widehat{\gamma}(0,1)|/\widehat{\sigma} > \delta/\sqrt{k_n}$ for some $t \in [0,1]$, the probability that should be controlled is

$$\Pr\left\{T_n > c_\alpha^S \,\middle|\, \frac{1}{\widehat{\sigma}} \sup_{t\in[0,1]} \{t|\widehat{\gamma}(0,t) - \widehat{\gamma}(0,1)|\} > \delta/\sqrt{k_n}\right\}, \tag{3}$$

and not $\Pr\left\{T_n > c_\alpha^S\right\}$ as in (2). Of course, the conditioning event in the above probability may also be written as $\{T_n > \delta\}$. We prefer to write it as in (3) to emphasize the 'formalized' eye-balling rationale of the conditioning.

Remark 1 To appreciate the difference between the conditional and unconditional approach, consider trajectories $X_1^{(b)}, \ldots, X_n^{(b)}$ $(b = 1, \ldots, B)$ for which (2) holds under $\mathcal{H}_0^S$. Then, for sufficiently large n, one expects to reject $\mathcal{H}_0^S$ in the unconditional procedure for $B\alpha$ of these trajectories. In contrast, the conditional approach considers

only those, say k, trajectories $\{X_1^{(b_i)}, \ldots, X_n^{(b_i)}\}_{i=1,\ldots,k}$ with visually apparent indications of change, i.e., those trajectories satisfying $\sup_{t\in[0,1]} \{t|\widehat{\gamma}(0,t) - \widehat{\gamma}(0,1)|\}/\widehat{\sigma} > \delta/\sqrt{k_n}$. If one rejects $\mathcal{H}_0^S$ for each of these k trajectories where the test statistic exceeds c_α^S, one cannot expect a rejection in (the desired number of) $k\alpha$ cases, but instead the number of rejections is much higher, since only trajectories with large fluctuations are considered in the first place.

Formally, the asymptotic rejection probability is given by the following

Theorem 1 *Suppose Assumptions 1 and 2 hold. Then, under $\mathcal{H}_0^S$, as $n \to \infty$,*

$$\Pr\left\{T_n > c_\alpha^S \,\middle|\, \frac{1}{\widehat{\sigma}} \sup_{t\in[0,1]} \{t|\widehat{\gamma}(0,t) - \widehat{\gamma}(0,1)|\} > \delta/\sqrt{k_n}\right\} \longrightarrow \begin{cases} 1, & \delta > c_\alpha^S, \\ \frac{\alpha}{f(\delta)}, & \delta \le c_\alpha^S, \end{cases} \tag{4}$$

where $f(x) = 2\sum_{n=1}^{\infty}(-1)^{n-1}\exp\{-2n^2x^2\} = \Pr\left\{\sup_{t\in[0,1]}|W(t) - tW(1)| \ge x\right\} \in [0,1]$.

As a simple application of the elementary conditional probability formula, the proof of Theorem 1 is almost trivial. The main insight afforded by Theorem 1 is that the textbook recommendation to look for 'any apparent changes in behavior' (Brockwell and Davis 2016, p. 12) should come with a warning that, if the formal structural break test is applied as usual (i.e., with the same critical value c_α^S suggested by the unconditional test), it rejects the null more often than it should. Moreover, the influence of the conditioning can be quantified up to the parameter δ. This is in contrast to the example of GoF tests mentioned in the Motivation. There, the act of looking at a histogram and spotting a resemblance with some parametric distribution is much harder to formalize and, hence, the impact on a subsequent GoF test much harder to quantify.

Specifically, Theorem 1 formalizes the intuition that if there is preliminary ('formalized' eye-balling) evidence in the data that a null hypothesis is false, then that null is more likely to be rejected, because the other cases—where there is no preliminary evidence—are not considered. Since the conditioning event in (4) may be equivalently written as $\{T_n > \delta\}$, it is clear that, in case $\delta > c_\alpha^S$, the true null is rejected not only asymptotically with probability one, but even almost surely in finite samples. Even for a less stringent testing condition with $\delta \le c_\alpha^S$ the conditional (asymptotic) rejection probability under the null, $\alpha/f(\delta)$, is larger than α. Only when $\delta = 0$, i.e., when there is no conditioning, is the desired type I error rate of α attained.

Nonetheless, if δ is known (which is typically not the case), there is a way to keep a desired confidence level α even in the conditional test. For a fixed $\delta > 0$, one simply chooses $\alpha^* = \alpha f(\delta) \in (0, \alpha)$. Then, by (4),

$$\Pr\left\{T_n > c_{\alpha^*}^S \,\middle|\, \frac{1}{\widehat{\sigma}} \sup_{t\in[0,1]} \{t|\widehat{\gamma}(0,t) - \widehat{\gamma}(0,1)|\} > \delta/\sqrt{k_n}\right\} \longrightarrow \alpha^*/f(\delta) = \alpha, \qquad n \to \infty.$$

This means that when the test is only applied if the data provide preliminary evidence for a structural break, the resulting bias can be corrected by suitably lowering the

confidence level of the critical value. In spirit, this is similar to the usual pre-testing strategy of (suitably) lowering the significance level that avoids inflating type I error (Giles and Giles 1993).

Remark 2 Theorem 1 only investigates the behavior of the test under the null and shows it to be oversized. In line with the well-known tradeoff between type I- and type II-error, this suggests that the test has higher power under the alternative. We refrain from formally investigating the (local) power of the test, as it is does not hold size and, hence, is not a valid statistical test.

2.2 Monitoring parameter changes

Let $X_1, \ldots, X_n$ denote the variables in the *training period* that are available prior to monitoring. The goal in monitoring is to detect breaks in some parameter as new observations $X_{n+1}, X_{n+2}, \ldots$ become available, and to do so as quickly as possible. Formally, interest in monitoring centers on sequentially testing

$$\mathcal{H}_0^M \ : \ \gamma = \gamma_{n+1} = \gamma_{n+2} = \cdots$$

given the *non-contamination* assumption $\gamma = \gamma_1 = \cdots = \gamma_n$, which imposes structural stability in the training period. Of course, non-contamination can be tested using the methods of Sect. 2.1. We consider so-called *closed-end* procedures, where monitoring stops after $X_{n+1}, \ldots, X_{\lfloor nT \rfloor}$ $(1 < T < \infty)$ have been observed.

Let $\widehat{\gamma}(a, b)$ denote a γ-estimate based on $X_{\lfloor na \rfloor + 1}, \ldots, X_{\lfloor nb \rfloor}$ $(0 \leq a < b \leq T)$. Define $D^2[0, T]$ as the space of all $\mathbb{R}^2$-valued functions on $[0, T]$ that are right-continuous with left-hand limits in each component (Davidson 1994). The following condition is the analogue of Assumption 1 and has likewise been shown to hold for various estimators (see, e.g., Hoga and Wied (2017)).

Assumption 3 It holds that, as $n \to \infty$,

$$\sqrt{k_n} \begin{pmatrix} t[\widehat{\gamma}(0,t) - \gamma] \\ t_0[\widehat{\gamma}(t, t+t_0) - \gamma] \end{pmatrix} \xrightarrow{d} \sigma \begin{pmatrix} W(t) \\ W(t+t_0) - W(t) \end{pmatrix} \qquad \text{in } D^2[0, T - t_0],$$

where $\sigma > 0, t_0 > 0, T > \max\{t_0, 1\}$, and $\{W(t)\}_{t \in [0,T]}$ denotes a standard Brownian motion.

We base monitoring on the moving-sum detector

$$M_n(t) = \frac{1}{\widehat{\sigma}} \left| t_0 \sqrt{k_n} [\widehat{\gamma}(t, t+t_0) - \widehat{\gamma}(0,1)] \right|, \qquad t \in [1, T - t_0],$$

where the estimator $\widehat{\sigma}$ from Assumption 2 is typically calculated from the non-contaminated training data. The idea behind $M_n(t)$ is that large deviations between $\widehat{\gamma}(t, t+t_0)$ and the non-contaminated estimate $\widehat{\gamma}(0,1)$ indicate a structural change in the monitoring period. Again by the continuous mapping theorem and Slutzky's

lemma, it follows under Assumptions 2 and 3 that

$$\sup_{t\in[1,T-t_0]} M_n(t) \xrightarrow{d} \sup_{t\in[1,T-t_0]} |W(t+t_0) - W(t) - t_0 W(1)|\,.$$

Thus, we reject $\mathcal{H}_0^M$ at significance level $\alpha \in (0,1)$ as soon as $M_n(t) > c_\alpha^M$ for some $t > 1$, where c_α^M is implicitly defined by

$$\Pr\left\{\sup_{t\in[1,T-t_0]} |W(t+t_0) - W(t) - t_0 W(1)| > c_\alpha^M\right\} = \alpha.$$

Hence, one controls the asymptotic probability

$$\Pr\left\{\sup_{t\in[1,T-t_0]} M_n(t) > c_\alpha^M\right\} \longrightarrow \alpha, \qquad n \to \infty. \tag{5}$$

When deciding which parameters to monitor, one may choose those whose subsample estimates have exhibited some variation in the training period. This again leads us to consider monitoring for change *conditional* on $\sup_{t\in[0,1]}\{t|\widehat{\gamma}(0,t) - \widehat{\gamma}(0,1)|\}/\widehat{\sigma} > \delta/\sqrt{k_n}$ for some $\delta > 0$. As before, the parameter δ is unknown, depending on the sensitivity of the visual inspection by the practitioner. When monitoring *conditionally* on having observed some noticeable variation in the training period, one should control

$$\Pr\left\{\sup_{t\in[1,T-t_0]} M_n(t) > c_\alpha^M \,\middle|\, \frac{1}{\widehat{\sigma}}\sup_{t\in[0,1]}\{t|\widehat{\gamma}(0,t) - \widehat{\gamma}(0,1)|\} > \delta/\sqrt{k_n}\right\}. \tag{6}$$

If δ could actively be chosen in practice, it should not be chosen too large (larger than some critical value c_α^S), because this would already be evidence against $\gamma_1 = \cdots = \gamma_n$ (cf. (2)), violating the non-contamination assumption. However, the conditioning in (6) does not matter asymptotically for monitoring, as shown next.

Theorem 2 *Suppose Assumptions 2 and 3 hold. Then, under $\mathcal{H}_0^M$, as $n \to \infty$,*

$$\Pr\left\{\sup_{t\in[1,T-t_0]} M_n(t) > c_\alpha^M \,\middle|\, \frac{1}{\widehat{\sigma}}\sup_{t\in[0,1]}\{t|\widehat{\gamma}(0,t) - \widehat{\gamma}(0,1)|\} > \delta/\sqrt{k_n}\right\} \longrightarrow \alpha. \tag{7}$$

Theorem 2 shows that when the training data inspire a test of $\mathcal{H}_0^M$, the monitoring procedure can be applied as usual, i.e., with the same boundary c_α^M as in (5). This result is reminiscent of the intuition that, while one cannot (without modification at least) use the same data that inspired a hypothesis to test it, one can simply wait for fresh (out-of-sample) data to verify it. The limit in (7) would trivially obtain if the two events in the probability were independent. Yet, this is not the case, as $\widehat{\gamma}(0,1)$ is common to both events, and the underlying X_t's may be serially dependent across the training and monitoring period.

Remark 3 The conclusion of Theorem 2 does not depend on the type of detector used. For instance, using the expanding sum detector $E_n(t) = \frac{1}{\widehat{\sigma}}|(t-1)\sqrt{k_n}[\widehat{\gamma}(1,t) - \widehat{\gamma}(0,1)]|$ would—under a suitable analogue of Assumption 3—lead to the same result. We omit details for brevity.

Remark 4 Reconsider the set-up of Remark 1. Suppose Assumptions 2 and 3 hold for the (continued) simulated trajectories $X_1^{(b)}, \ldots, X_n^{(b)}, X_{n+1}^{(b)}, \ldots$ $(b = 1, \ldots, B)$. Then, in sufficiently large samples, one expects to reject $\mathcal{H}_0^M$ based on *unconditional* testing for roughly $\alpha\%$ of all trajectories; see (5). (This is as in the structural break setting of Sect. 2.1, where one also rejects $\mathcal{H}_0^S$ for $\alpha\%$ of all trajectories.) If monitoring is done *conditionally*, then one expects to reject $\mathcal{H}_0^M$ for $\alpha\%$ of the k trajectories satisfying the conditioning event. (This contrasts with the structural break tests, where—in the conditional setting—$\mathcal{H}_0^S$ is rejected for $\alpha/f(\min\{\delta,\alpha\})\%$ of said k trajectories.) Hence, while (5) and (7) suggest an equal number of type I errors asymptotically, the trajectories tested are different—in unconditional monitoring *all* trajectories are considered, whereas conditional monitoring only considers the *fraction*, where the condition is met.

Remark 5 The result illustrated in Remark 4 for the different conditional rejection probabilities in structural break testing and monitoring has some analogy with the following example. Suppose a coin comes up heads 9 out of 10 times. This then raises some suspicion of the fairness of the coin. The data-inspired hypothesis that the coin is unfair, is then more likely to be accepted as true when tested on the same observations. (This result is analogous to Theorem 1.) However, tossing the same coin again 10 times, the 'unfair coin' hypothesis—suggested by the first 10 throws—can be tested on fresh data as if no conditioning took place. (This is analogous to Theorem 2, where out-of-sample data become available for monitoring.) Of course, while fresh data is also used in testing $\mathcal{H}_0^M$, the monitoring situation is more complex than the coin flip example. First, data may be serially dependent in Theorem 2. Second, $\widehat{\gamma}(0,1)$ appears in both the detector $M_n(t)$ and the conditioning event in (6).

3 Empirical application

Here, we illustrate the practical implications of Theorems 1 and 2 . We do so using Bitcoin log-returns $X_1, \ldots, X_n$ from 01/01/2016 to 31/12/2019, giving $n = 1,461$ observations.[1] Böhme et al. (2015) provide a comprehensive review of the crypto-currency. Due to the rising popularity of crypto-currencies, much research effort has been devoted to studying Bitcoin, which represents the largest market share among all crypto-currencies. For instance, Urquhart (2016) investigates the efficiency of the Bitcoin market using, among others, a classical Ljung–Box test. Most tests (including Ljung–Box tests) rely on the absence of structural breaks for their validity. However, as Bitcoins represent a new asset class, ex ante knowledge of the parameters that may change (e.g., mean, variance, higher order moments, tail index, autocorrelations, etc.) is hard to justify. Testing for change in all conceivable parameters invariably inflates

[1] The data were downloaded from *finance.yahoo.com* (Ticker Symbol: *BTC-USD*).

type I errors, due to multiple testing issues. So it seems natural to only test for breaks in parameters that have fluctuated noticeably in the data. Hence, in the following, we test for breaks in parameters, where 'formalized' eye-balling—as described in Sect. 2.1—indicates a possible change.

We exemplarily consider the first two moments as the most important parameters determining location and scale. On stretches of stationarity in the data, Assumptions 1–3 are then likely to be satisfied by the Bitcoin log-returns. This is because GARCH-type volatility models have been successfully used in modeling Bitcoin returns (Cheah and Fry 2015). Carrasco and Chen (2002) show that several GARCH models are mixing with rates implying suitable functional central limit theorems in Assumptions 1 and 3 to hold (Herrndorf 1985). Likewise the mixing conditions are sufficient for consistent estimation of the long-run variance σ^2 in Assumption 2 (De Jong and Davidson 2000).

Fix the significance level at $\alpha = 0.05$. In the 'standard' eye-balling method, one would plot the series of log-returns, and look for preliminary evidence of parameter change. If such evidence is found, the series would then be subjected to a formal level-α test, without reflecting in the test that the data themselves suggested the hypothesis. Alternatively, we use the 'formalized' eye-balling approach and assume that visually inspecting the time series for abrupt changes inspires a subsequent test if and only if $\frac{\sqrt{n}}{\widehat{\sigma}} \sup_{t\in[0,1]} \{t|\widehat{\gamma}(0,t) - \widehat{\gamma}(0,1)|\}$ exceeds δ. Of course, δ is unknown in practice, but for illustrative purposes we assume here that it equals the 20%-critical value, i.e., $\delta = c^S_{\alpha=0.2} = 1.073$. Thus, we apply a test to one of the two parameters—first and second moments—if the corresponding value of T_n is larger than δ. Without any modification of the significance level α, this would yield a test of level $\alpha/f(\delta) = \alpha/0.2 = 25\%$ by Theorem 1. By the same theorem, a (conditional) level-α test is however obtained, if we reject when $T_n > c^S_{\alpha^*} = 1.628$ for $\alpha^* = \alpha f(\delta) = 0.2\alpha = 0.01$.

Figure 1 displays the price process and the log-returns of Bitcoin. The prices seem to indicate returns with positive mean until the peak on December 16, 2017, and a negative mean in the year thereafter. This strongly suggests the need for a more formal test of a constant mean. Define $S^M_n(t) = \frac{\sqrt{n}}{\widehat{\sigma}^M} t|\widehat{\gamma}^M(0,t) - \widehat{\gamma}^M(0,1)|$, where $\widehat{\gamma}^M(0,t)$ denotes the sample mean of $X_1, \ldots, X_{\lfloor nt \rfloor}$, and $\widehat{\sigma}^M$ is a HAC estimator with Bartlett kernel and bandwidth $\lfloor \log n \rfloor$. As the test statistic $T^M_n = \sup_{t\in[0,1]} S^M_n(t) = 1.578$ is larger than the critical value $c^S_{\alpha=0.05} = 1.358$, the null of a constant mean is rejected. However, this test incurs a not-accounted-for bias, because the mean break hypothesis was suggested by the same data that were used for testing.

To account for this bias, we apply the 'formalized' eye-balling technique next. The test statistic for a mean change is larger than δ ($T^M_n = 1.578 > 1.073 = \delta$). Conditional on this result, a level-α test rejects if $T^M_n > 1.628$. Hence, we do not reject the constant mean hypothesis at a 5%-significance level using the *conditional* test.

The plot of $t \mapsto S^M_n(t)$ in Fig. 1 graphically illustrates the conflicting results. The red dotted lines indicate the 'incorrect' critical value $c^S_{\alpha=0.05} = 1.358$ and the 'correct' $c^S_{\alpha=0.01} = 1.628$. While the 'incorrect' value is exceeded by $S^M_n(t)$, the 'correct' value is not. The additional evidence required to exceed the 'correct' critical value can be seen as a compensation for the fact that the data themselves indicated the mean break hypothesis.

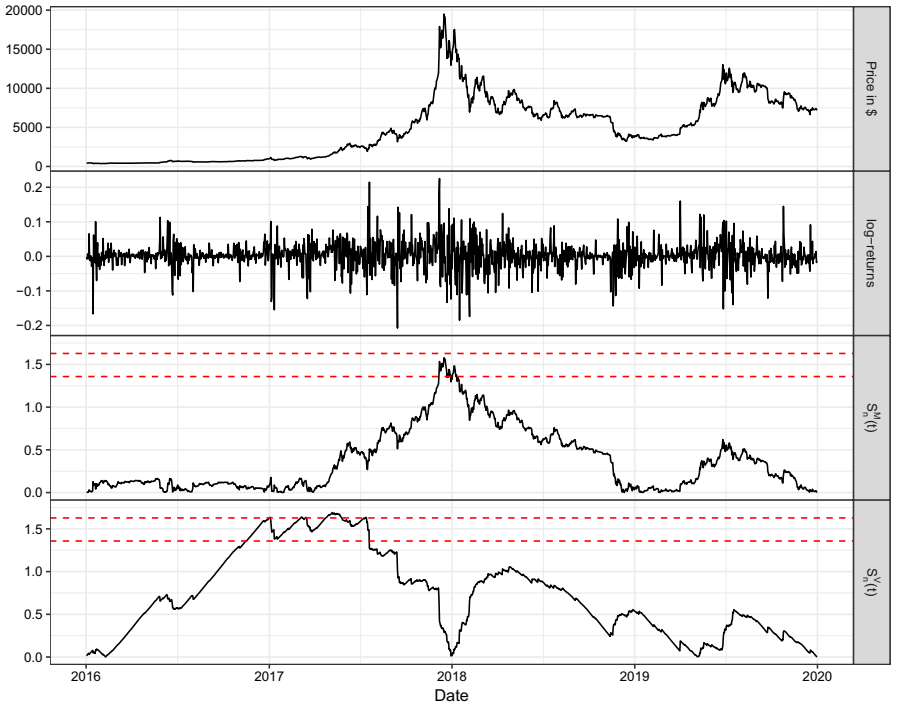

Fig. 1 From top to bottom: Bitcoin prices in US $, log-returns, plots of $t \mapsto S_n^M(t)$ and $t \mapsto S_n^V(t)$. The dashed red lines in the bottom two panels indicate the 1%- and 5%-critical values $c^S_{\alpha=0.01} = 1.628$ and $c^S_{\alpha=0.05} = 1.358$

The valid conditional test did not reject the constant mean hypothesis. Nonetheless, the persistence in the price process observed in Fig. 1 indicates the possibility of mean changes. This suggests that it may be useful to monitor for breaks in the mean starting in 2020. For instance, in a risk management context, it is particularly important to detect downward breaks in the mean to avoid losses. The implication of Theorem 2 is that such a monitoring procedure can be applied as if it had not been Fig. 1 that suggested the hypothesis.

As the conditional test did not reject the null of a constant mean, we may validly apply a test for a change in the variance by testing for the constancy of second moments. Define $S_n^V(t) = \frac{\sqrt{n}}{\widehat{\sigma}^V} t|\widehat{\gamma}^V(0,t) - \widehat{\gamma}^M(0,1)|$, where $\widehat{\gamma}^V(0,t)$ denotes the sample mean of $X_1^2, \ldots, X_{\lfloor nt \rfloor}^2$, and $\widehat{\sigma}^V$ again denotes a HAC estimator (with Bartlett kernel and bandwidth $\lfloor \log n \rfloor$) for the asymptotic variance of $\widehat{\gamma}^V(0,1)$. The test statistic for a variance change is once more larger than δ ($T_n^V = \sup_{t\in[0,1]} S_n^V(t) = 1.692 > 1.073 = \delta$). Conditional on this result, a level-α test rejects the null of a constant variance, as $T_n^V = 1.692 > 1.628$. Of course, in this case the 'naive' test also would have led to a rejection (because $T_n^V = 1.692 > 1.358 = c^S_{\alpha=0.05}$). The plot of $t \mapsto S_n^V(t)$ in Fig. 1 illustrates the result. It also allows to date the (most prominent) break somewhere around the first half of 2017, where the largest values of $S_n^V(t)$

are attained. Of course, as the variance is not even constant in the training period, monitoring (using Theorem 2) should not be carried out, due to structural change contaminating the training period.

The evidence for a break in the variance of Bitcoin returns suggests that *either* statistical analyses (such as Ljung–Box tests used to assess market efficiency) need to be robust to unconditional heteroscedasticity *or* the analyses have to be restricted to break-free subsamples.

4 Conclusion

More often than not, hypotheses are generated by data. While, in general, fresh data is desirable to validly verify a hypothesis, in some applications hypotheses need to be tested on the same data that generated it (e.g., the structural break hypothesis for the Bitcoin returns from 2016 to 2019). In situations like these, the bias of having looked at the data before hypotheses are formulated can frequently not be corrected, e.g., in goodness-of-fit testing. We show in this note that a correction is theoretically possible for structural break tests, if the critical value is suitably increased—with the increase depending on a single unknown constant. This provides one further reason for the use of large critical values or, equivalently, small significance levels, that has also been advocated elsewhere (Benjamin 2018). Furthermore, this shows that the textbook recommendation to visually inspect the data for breaks should carry a warning that subsequent formal structural break tests need to take into account that the break hypothesis was suggested by the data. By contrast, when monitoring parameters, 'hypotheses' generated from the training data can be tested without any correction.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00362-021-01233-4.

Funding Open Access funding enabled and organized by Projekt DEAL.

Appendix

Proof of Theorem 1 We use the elementary conditional probability formula to write

$$\Pr\left\{T_n > c_\alpha^S \;\middle|\; \frac{1}{\widehat{\sigma}} \sup_{t\in[0,1]} \{t|\widehat{\gamma}(0,t) - \widehat{\gamma}(0,1)|\} > \delta/\sqrt{k_n}\right\}$$

$$
\begin{aligned}
&= \Pr\left\{ T_n > c_\alpha^S \mid T_n > \delta \right\} \\
&= \Pr\left\{ T_n > c_\alpha^S,\ T_n > \delta \right\} / \Pr\left\{ T_n > \delta \right\} \\
&= \Pr\left\{ T_n > \max\{c_\alpha^S,\ \delta\} \right\} / \Pr\left\{ T_n > \delta \right\}.
\end{aligned}
$$

The conclusion now follows from (1) and Proposition 12.3.4 of Dudley (2004). □

Proof of Theorem 2 From the continuous mapping theorem, Slutzky's lemma and Assumptions 2 and 3, we obtain, as $n \to \infty$,

$$
\begin{aligned}
&\Pr\left\{ \sup_{t\in[1,T-t_0]} M_n(t) > c_\alpha^M \;\middle|\; \frac{1}{\widehat{\sigma}} \sup_{t\in[0,1]} \{t|\widehat{\gamma}(0,t) - \widehat{\gamma}(0,1)|\} > \delta/\sqrt{k_n} \right\} \\
&= \frac{\Pr\left\{ \sup_{t\in[1,T-t_0]} \frac{1}{\widehat{\sigma}} \left|t_0\sqrt{k_n}[\widehat{\gamma}(t,t+t_0) - \widehat{\gamma}(0,1)]\right| > c_\alpha^M,\ \frac{\sqrt{k_n}}{\widehat{\sigma}} \sup_{t\in[0,1]} |t[\widehat{\gamma}(0,t) - \widehat{\gamma}(0,1)]| > \delta \right\}}{\Pr\left\{ \frac{\sqrt{k_n}}{\widehat{\sigma}} \sup_{t\in[0,1]} |t[\widehat{\gamma}(0,t) - \widehat{\gamma}(0,1)]| > \delta \right\}} \\
&\longrightarrow \frac{\Pr\left\{ \sup_{t\in[1,T-t_0]} |W(t+t_0) - W(t) - t_0 W(1)| > c_\alpha^M,\ \sup_{t\in[0,1]} |W(t) - tW(1)| > \delta \right\}}{\Pr\left\{ \sup_{t\in[0,1]} |W(t) - tW(1)| > \delta \right\}}.
\end{aligned}
\tag{A.1}
$$

The processes $\{W(t+t_0) - W(t) - t_0 W(1)\}_{t\in[1,T-t_0]}$ and $\{W(s) - sW(1)\}_{s\in[0,1]}$ are independent, because they are both Gaussian and uncorrelated:

$$
\begin{aligned}
&\mathrm{E}\left\{ \left[W(t+t_0) - W(t) - t_0 W(1)\right]\left[W(s) - sW(1)\right] \right\} \\
&\quad = \min\{t+t_0, s\} - s\min\{t+t_0, 1\} - \min\{t, s\} + s\min\{t, 1\} - t_0 \min\{1, s\} + t_0 s \min\{1, 1\} \\
&\quad = s - s - s + s - t_0 s + t_0 s = 0.
\end{aligned}
$$

Hence, the suprema in the numerator of (A.1) are independent, and the ratio in (A.1) reduces to

$$
\Pr\left\{ \sup_{t\in[1,T-t_0]} |W(t+t_0) - W(t) - t_0 W(1)| > c_\alpha^M \right\} = \alpha
$$

by definition of c_α^M. The conclusion follows. □

References

Baltagi BH, Kao C, Na S (2013) Testing for cross-sectional dependence in a panel factor model using the wild bootstrap F test. Stat Pap 54(4):1067–1094

Benjamin DJ et al (2018) Redefine statistical significance. Nat Hum Behav 2:6–10

Böhme R, Christin N, Edelman B, Moore T (2015) Bitcoin: economics, technology, and governance. J Econ Perspect 29:213–238

Brockwell PJ, Davis RA (2016) Introduction to time series and forecasting, 3rd edn. Springer, New York

Carrasco M, Chen X (2002) Mixing and moment properties of various GARCH and stochastic volatility models. Econ Theory 18:17–39

Cheah E-T, Fry J (2015) Speculative bubbles in bitcoin markets? An empirical investigation into the fundamental value of bitcoin. Econ Lett 130:32–36
Davidson J (1994) Stochastic limit theory. Oxford University Press, Oxford
De Jong RM, Davidson J (2000) Consistency of Kernel estimators of heteroscedastic and autocorrelated covariance matrices. Econometrica 68(2):407–423
Dehling H, Vogel D, Wendler M, Wied D (2017) Testing for changes in Kendall's Tau. Econ Theory 33:1352–1386
Demetrescu M, Hanck C (2013) Nonlinear IV panel unit root testing under structural breaks in the error variance. Stat Pap 54(4):1043–1066
Dudley RM (2004) Real analysis and probability. Cambridge University Press, Cambridge
Gelman A, Loken E (2014) The statistical crisis in science. Am Sci 102:460–465
Giles JA, Giles DEA (1993) Pre-test estimation and testing in econometrics: recent developments. J Econ Surv 7:145–197
Harvey DI, Leybourne SJ, Taylor AMR (2013) Testing for unit roots in the possible presence of multiple trend breaks using minimum Dickey–Fuller statistics. J Econ 177:265–284
Herrndorf N (1985) A functional central limit theorem for strongly mixing sequences of random variables. Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 69(4):541–550
Hoga Y (2017a) Change point tests for the tail index of β-mixing random variables. Econ Theory 33(4):915–954
Hoga Y (2017b) Testing for changes in (extreme) VaR. Econ J 20(1):23–51
Hoga Y (2018) A structural break test for extremal dependence in β-mixing random vectors. Biometrika 105:627–643
Hoga Y, Wied D (2017) Sequential monitoring of the tail behavior of dependent data. J Stat Plan Inference 182:29–49
Inclan C, Tiao GC (1994) Use of cumulative sums of squares for retrospective detection of changes of variance. J Am Stat Assoc 89:913–923
Kerr NL (1998) HARKing: hypothesizing after the results are known. Perspect Soc Psychol Rev 2(3):196–217
Quintos C, Fan Z, Phillips PCB (2001) Structural change tests in tail behaviour and the Asian crisis. Rev Econ Stud 68:633–663
Ruppert D, Matteson DS (2015) Statistics and data analysis for financial engineering, 2nd edn. Springer, New York
Selvin HS, Stuart A (1966) Data-dredging procedures in survey analysis. Am Stat 20:20–23
Simmons JP, Nelson LD, Simonsohn U (2011) False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol Sci 22(11):1359–1366
Urquhart A (2016) The inefficiency of bitcoin. Econ Lett 148:80–82
Vogelsang TJ (1998) Testing for a shift in mean without having to estimate serial-correlation parameters. J Bus Econ Stat 16:73–80
Wied D, Arnold M, Bissantz N, Ziggel D (2012) A new fluctuation test for constant variances with applications to finance. Metrika 75:1111–1127
Wied D, Krämer W, Dehling H (2012) Testing for a change in correlation at an unknown point in time using an extended functional delta method. Econ Theory 28:570–589
Xu K-L (2015) Testing for structural change under non-stationary variances. Econ J 18:274–305