Statistical Methods for Large Financial Data Sets: Essays on Monitoring Cointegration and Distributional Reference Class Forecasting

Inauguraldissertation
zur
Erlangung des Doktorgrades
der
Wirtschafts- und Sozialwissenschaftlichen Fakultät
der
Universität zu Köln

2024

vorgelegt
von

M.Sc. Etienne Theising

aus

Rheine

Referent:	Prof. Dr. Dominik Wied
Korreferent:	Prof. Dr. Jörg Breitung
Tag der Promotion:	13. August 2024

Acknowledgements

First and foremost, I want to express my gratitude to my main doctoral supervisor Prof. Dr. Dominik Wied. He has already supported me during my bachelor and master studies, has given me the opportunity to write this doctoral dissertation and has provided friendly guidance and helpful comments whenever they were necessary.

I am further thankful to Prof. Dr. Jörg Breitung for accepting to be my secondary supervisor.

Moreover, I am grateful to Prof. Dr. Daniel Ziggel for bringing the topic of reference class selection to my attention and offering valuable discussions and help on my second project.

Further, I thank all colleagues of the Institute of Econometrics and Statistics at the University of Cologne. There has always been a positive atmosphere and encouraging spirit among all members of the institute and I enjoyed the time there with plenty of discussions.

Lastly, I want to thank my parents for their constant support during my educational years and, most importantly, I express my deepest gratitude to Sylvie, Rieke and Joris for their continuous patience and belief in me.

Without all of the above mentioned, the endeavor to write this doctoral dissertation would not have been possible.

Contents

List of Tables

List of Figures

Introduction

Ever since the digital revolution and related emerging technologies data have become increasingly relevant. This phenomenon stretches out to scientific data as well (Diebold, 2003) and the increase in data availability coined the term "Big Data", initially defined as data sets that are too large to be handled by existing machinery and analysis tools (Fan and Bifet, 2013), but now loosely understood as a vast amount of data in the general public and academia while lacking a general definition (see, e.g., Kitchin and McArdle, 2016). Large data sets provide research opportunities in practice, for instance, by new data insights, making applications of limit theorems feasible or improving parameter estimation due to larger sample sizes. Simultaneously, large data sets call for novel analysis tools, e.g., to handle the data computationally or to select subsets from the data in a sophisticated way for subsequent research.

This doctoral dissertation consists of three essays dealing with methods suitable for application on data sets that can be considered large within the econometrics and finance literature.[1] In the respective methods covered here, data quantity has two different origins. On the one hand, the theory developed in Chapter 1 is suitable for data sets with a short cross-section but large time series component. The latter is exploited by a central limit theorem and used to monitor the structural stability of a data model, namely systems of cointegrating regressions, with a sequential hypothesis test. The monitoring is investigated for panels with up to 30 cross-sections and between 500 and 1,000 time periods on three variables where asymptotic results generally allow for more time periods. On the other hand, the approaches to use reference classes established in Chapters 2 and 3 are suited for data sets that are at least moderate-sized in time dimension and large in cross-section dimension. Here, the amount of data enables a general approach to (distributional) forecasting. Moreover, the theory simultaneously offers a method for data reduction in order to select a relevant subset from the data. Reference classes are selected, i.a., according to recency and, thus, time series information are only relevant

[1] Compared to other scientific areas we deal with rather *small* data sets, for example, the CERN released roughly 800 terabytes of data in December 2023 that were collected by Large Hadron Collider experiments in 2011 and 2012 (see https://home.cern/news/news/experiments/lhcb-experiment-releases-all-its-run-1-proton-proton-data).

if the observations are close enough to the forecast case. Reference class forecasting is examined on a data set with 21,808 cross-sections and 70 time periods on 28 variables while the method can handle even larger data sets in general due to its computational simplicity. The remainder of the introduction includes a more detailed description of the chapters.

While the first two chapters are joint work with Dominik Wied (c.f. Theising and Wied, 2023, published in *Econometris and Statistics*) and Dominik Wied as well as Daniel Ziggel (c.f. Theising et al., 2023, published in *Journal of Forecasting*), the last chapter is based on a single-author paper (c.f. the *working paper* Theising, 2024).

Chapter 1 covers new residual-based monitoring statistics for structural changes in systems of cointegrating relationships. Reviewing the structural stability of time series models is crucial and the approach is based on parameter estimation over a calibration period as in Chu et al. (1996) while the date of the potential breakpoint does not need to be known a priori. In case of homogenous systems and cross-sectional independence the pooled fully modified OLS estimator (PFM-OLS, Phillips and Hansen, 1990) takes into account the effects of error serial correlation and regressor endogeneity. Cross-sectional dependence is allowed for by using the pooled fully modified GLS estimator (PFM-GLS) for homogenous systems and the fully modified SUR estimator (FM-SUR, Moon, 1999) for inhomogenous systems. The monitoring procedure for systems of cointegrating regressions is an extension of the single equation monitoring by Wagner and Wied (2017). Three detectors for each estimation technique are proposed and the limiting distributions of the monitoring statistics under the null hypothesis are derived. The monitoring statistics show decent behaviour under the null hypothesis with controlled rejection probabilities and power against two alternatives for different data generating processes. An empirical application investigates deviations from the triangular arbitrage parity condition for three bivariate panels of exchange rate triplets including Bitcoin. The procedures detect breakpoints between May and August 2014 and between January and May 2015 indicating an instability in arbitrage parities. Following this, a promising portfolio trading strategy based on the breakdates is constructed and compared to a simple buy-and-hold strategy.

A similar topic, but limited to PFM-OLS detectors, was covered in my master thesis “Monitoring Cointegration in a System of Homogeneous Cointegrating Regressions” (Theising, 2018). In particular, Lemma 3 is already stated there but the simulation results using PFM-OLS detectors presented here are new. During my doctoral studies, I made the following contributions to Chapter 1: I wrote and revised the main body of text along with the appendix. Further, I proposed to use PFM-GLS and FM-SUR estimators to relax the assumptions allowing for cross-sectional dependence and heterogeneous parameters and worked out the technical details thereof including nontrivial theoretical results. All R code used throughout the project was authored by me. I worked out how to simulate critical values, implemented this and conducted the simulation study. For the application, I collected the necessary data and carried out both the empirical application on exchange rate triplets and the portfolio trading strategy.

Chapter 2 presents a general method to handle forecasts exposed to behavioural bias by finding appropriate outside views. In this case, corporate sales suffer from low predictability (see Chan et al., 2003) and forecasts of analysts are often based on heuristics and were empirically shown to be biased as well as overoptimistic (see, e.g., Tversky and Kahneman, 1974; Kahneman and Tversky, 1973). The idea is to construct reference classes, similar to peer groups, for each examined company separately. Elements of the reference class should be similar to an examined firm in terms of a specific covariate, here called reference variable, that serves implicitly as a predictor. The reference classes then offer statistical information by the empirical distribution of sales growth of these similar firms which can be used to identify bias if the forecasts seem extreme compared to the distribution within the reference classes (Kahneman and Tversky, 1979). Reference classes are regarded to be optimal if the sales distributions within them match the distributions of realized future sales growth as closely as possible. Additionally, the empirical distribution within the reference class can be seen as a distributional forecast and the respective quality is measured by applying goodness-of-fit tests on the estimated probability integral transform values and by comparing the predicted quantiles thereof with a novel measure Δ_q. The method is out-of-sample backtested on a data set consisting of 21,808 US firms over the time period from 1950 to 2019. It appears that in particular the past operating margins are good reference variables for the reference class selection of future sales growth. A case study compares the outside view of constructed reference classes, issued as distributional forecasts, with actual analysts' forecasts and emphasizes the relevance of the approach in practice.

Throughout Chapter 2, I conceptualized the theoretical framework and obtained and preprocessed the Compustat, CRSP and inflation rate data sets. I combined them and described the resulting data set. All R code for this project was written by me and I conducted the backtest and empirical application. Further, I contributed to the description of backtest and application results and wrote parts of the manuscript, including the introduction and conclusion.

Chapter 3 extends the approach to reference class selection in distributional forecasting from Chapter 2 to using several co-variates as reference variables in view of the general reference class problem in statistics (Venn, 1888; Reichenbach, 1949). A general framework for including multiple variables is described and the focus lies on distributional forecasts that arise from reference classes. Rank-based algorithms are proposed for reference class selection including an optional preprocessing data dimension reduction via principal components analysis. Here, ranks robustify against skewness and outlier effects in the underlying data. A review of methods to evaluate distributional forecasts, such as probabilty integral transform values (proposed by Dawid, 1984; Diebold et al., 1998), statistical goodness-of-fit tests and proper scoring rules (Gneiting and Raftery, 2007), places the measure Δ_{q} from Chapter 2 in the literature. As an illustrative application, forecasting corporate sales growth rates is revisited to allow for a meaningful evaluation of the results based on multiple reference variables compared to the findings in Chapter 2 on the same data set following the prequential principle (Dawid, 1984). Again, probability integral transform values determine Δ_{q} which ranks the distributional forecast capability of different reference variable sets and algorithms in a backtest using a forward selection and a brute force approach on selected reference variable subsets. Particularly, algorithms on dimension reduced variables perform well using contemporaneous balance sheet and financial market parameters along with past sales growth rates and past operating margin changes. Finally, the practical use of the method is illustrated by predictions of interval probabilities, a comparison between historic distributional sales growth forecasts and realized sales growth, and two comparisons of actual analysts' estimates and distributional one-year ahead sales growth forecasts.

Chapter 1.

Monitoring Cointegration in Systems of Cointegrating Relationships

1.1. Introduction

This chapter proposes residual based self-normalized monitoring procedures for structural change in a system of homogeneous cointegrating regressions. Such procedures might be useful to detect deviations from stable economic relationships, e.g., macroeconomic equations for housing prices or financial equations for exchange rates. There is recent empirical evidence that such relationships might collapse (Anundsen, 2015, for the subprime bubble, Reynolds et al., 2021, and Reynolds et al., 2018, for cryptocurrencies) and we provide a methodological contribution to formally detect such collapses as early as possible. This is relevant from an economic point of view, but also for potential subsequent econometric analyses (see, e.g., Arsova and Örsal, 2021). As we assume that the number of cross-sections N is small and the number of time periods T is large, our procedure is typically most relevant for financial data.

Our asymptotically valid panel data method is an extension of the single equation monitoring procedure of Wagner and Wied (2017). Specifically, on the one hand, we assume homogeneous parameters and cross-sectional independent and identically distributed errors. On the other hand, we discuss extensions to the cases of heterogenous parameters and cross-sectional dependence. Our procedures are consistent if the cointegrating relationship turns to a spurious regression or if there is a break in the trend and/or slope parameters. The date of the potential change points does not need to be known a priori. Our monitoring procedures require parameter estimates and a monitoring statistic. We follow the ideas of Chu et al. (1996) and base the parameter estimates on a break-free (or assumed to be break-free) calibration period as a fraction of the whole sample size. The monitoring procedures use residuals calculated from these parameter estimates to calculate cointegration test statistics over expanding windows. Since the limiting distributions of our test statistics depend on the fraction of the calibration period, we in fact propose "closed-end" monitoring procedures, i.e., the monitoring horizon has to be specified beforehand.

We use the pooled fully modified OLS (PFM-OLS) estimator by Phillips and Moon (1999) to obtain nuisance parameter free null limiting distributions of the monitoring statistics[1] and, newly, a pooled fully modified feasible GLS (PFM-GLS) estimator as well as the fully modified SUR (FM-SUR) estimator by Moon (1999) in case of cross-sectionally dependent cointegrating regressions. The limiting distributions also depend on the choice of deterministic regressors as well as the number of I(1)-regressors. Our monitoring statistics are based on the properly scaled partial sum process of FM-OLS-type residuals and are inspired by the statistics in Wagner and Wied (2017) which are based on the statistic of the Shin (1994) test. We analyze our approach with respect to different transformations from a multivariate partial sum process to a scalar test statistic.

A simulation study assesses the performance of the PFM-OLS procedure in terms of rejection probabilities under the null hypothesis as well as power and detection delays under various alternatives, including influence of regressor endogeneity and serial correlation, sample size and fraction of the calibration period m. Under the null hypothesis the procedures work well in terms of null rejection probabilities close to the chosen significance level. We further investigate how the PFM-GLS and FM-SUR procedures work under cross-sectional dependence assumptions. For a variety of alternatives we investigate both power and detection times, which serve as natural estimates of potential breakpoints. Finally, we provide simulations which indicate that, in terms of null rejection probabilities, it is advisable to choose a monitoring period as long as possible such that the calibration period is as little a fraction of the whole period as possible.

We provide a test for stability in bivariate systems of homogeneous cointegrating relationships in triangular arbitrage parities for logarithmic exchange rate triplets including Bitcoin as an illustrative application example. We apply the procedures to a stochastic variant of the aforementioned parity arising from no-arbitrage assumptions between triplets of currency exchange rates such that there is no profit in instantaneous circular transactions. We assume that violations of triangular arbitrage parities under normal market conditions are stationary and a turn to non-stationary deviations is a sign of mispricing not due to financial frictions – also referred to as financial market dislocation. Reynolds et al. (2021) find empirical evidence of such mispricing in currency triplets

[1] Partially covered in my master thesis (Theising, 2018).

including Bitcoin using the Wagner and Wied (2017) monitoring for single equation cointegrating relationships and use their results for a currency portfolio strategy. Our sample ranges from 1 May 2013 until 31 December 2015, while the calibration period stretches until 8 November 2013, assuming a break free calibration period due to stable Bitcoin prices. The monitoring statistics indicate structural change between May and August 2014 and between January and May 2015 for some pairs of exchange rate triplets. Important dates during monitoring and prior to the detected breaks are the ending of the cap on euro-swiss franc exchange rates by the Swiss National Bank in January 2015 and, in February 2014, the closing of Mt. Gox, a Japanese Bitcoin exchange where 70% of all tradings took place up to its closing (Decker and Wattenhofer, 2014), which in turn resulted in the loss of $850,000$ Bitcoin with a total value of 473 million USD at that time (Fink and Johann, 2014). Reynolds et al. (2021) do not account for testing several cointegrating relationships at a time, in contrast, these monitoring procedures do. We apply our results to construct a portfolio trading strategy using the detected breaks as a sign of currency market instabilities.

Section 1.2 presents the model, the assumptions as well as the monitoring statistics. Section 1.3 presents the results of the simulation study, whilst Section 1.4 is dedicated to the application. Section 1.5 briefly summarizes and concludes. Three appendices follow the main text: Appendix 1.A contains all proofs, Appendix 1.B describes the simulation of critical values and Appendix 1.C shows additional results on error covariances of the application and selected simulation cases.

1.2. Monitoring Systems of Cointegrating Regressions

We consider monitoring the structure in a system of N cointegrating relationships (also referred to as cointegrating regressions or cointegrating equations) with a potential change point

$$y_{n,t} = \begin{cases} D_t'\theta_{D,n} + X_{n,t}'\theta_{X,n} + u_{n,t} & , t = 1, \dots, [rT], \\ D_t'\theta_{D,1,n} + X_{n,t}'\theta_{X,1,n} + u_{n,t} & , t = [rT]+1, \dots, T, \end{cases} \tag{1.1}$$

and

$$\Delta X_{n,t} = v_{n,t}, \quad t = 1, \dots, T, \tag{1.2}$$

for $n = 1, \dots, N$, i.e., the setting from Wagner and Wied (2017) extended to multiple cointegrating equations. Throughout the paper, we assume a break-free calibration period

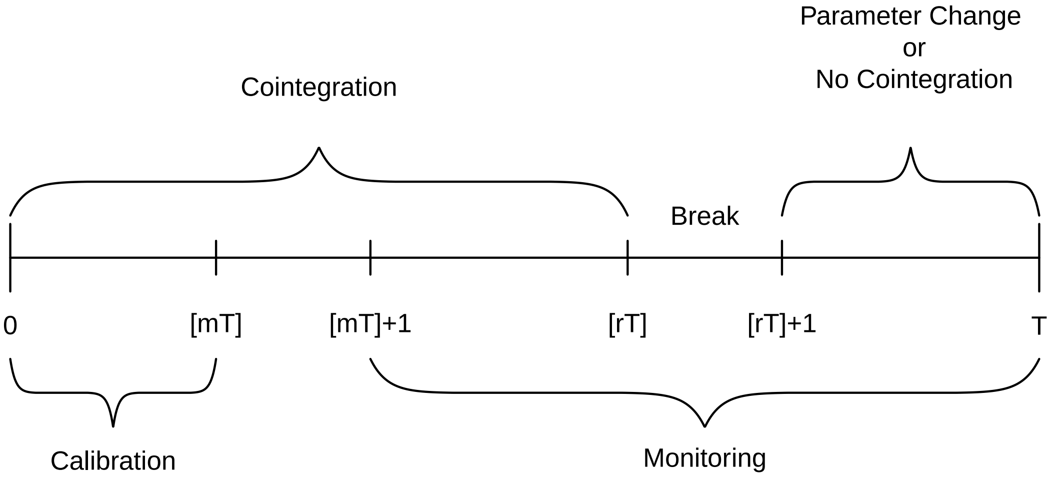

Figure 1.1.: Illustration of the monitoring procedure

of length $[mT]$ at the sample beginning and consider the case of small N and large T, i.e., for asymptotics N is fixed and $T \to \infty$. $y_{n,t}$ is scalar, $D_t \in \mathbb{R}^p$ is the deterministic trend function, $X_{n,t}$ is a non-cointegrated k-dimensional random vector of I(1) regressors, $u_{n,t}$ is a zero mean error process and $0 < m \leq r < 1$. We allow endogeneous regressors and serial correlation in the zero mean errors $v_{n,t} = [v_{n,t,1}, \dots, v_{n,t,k}]'$ of $X_{n,t}$ as well as correlation across k and n. Let $\theta_n = [\theta'_{D,n}, \theta_{X,n}]'$ and $\theta_{1,n} = [\theta'_{D,1,n}, \theta_{X,1,n}]'$.

We test the following pair of hypotheses:

$$H_0 : \begin{cases} \theta_n = \theta_{1,n} \text{ for all } m \leq r < 1, n = 1, \dots, N, \text{ and} \\ \{u_{n,t}\}_{t=1,\dots,T} \text{ is I(0) for all } n = 1, \dots, N \end{cases} \tag{1.3}$$

and

$$H_1 : \begin{cases} \theta_n \neq \theta_{1,n} \text{ for some } m \leq r < 1, n \in \{1, \dots, N\} \textbf{ or} \\ \{u_{n,t}\}_{t=1,\dots,[rT]} \text{ is I(0) and } \{u_{n,t}\}_{t=[rT]+1,\dots,T} \text{ is I(1)} \\ \text{for some } m \leq r < 1, n \in \{1, \dots, N\} \end{cases} \tag{1.4}$$

There is no structural change under the null hypothesis, i.e., $\theta_{D,n} = \theta_{D,1,n}$, $\theta_{X,n} = \theta_{X,1,n}$ and $\{u_{n,t}\}_{t=1,\dots,T}$ is I(0) for all $n = 1, \dots, N$. Under the alternative either a change in the parameter occurs or cointegration turns to spurious regression in at least one cointegrating relationship at a sample fraction $[rT]$ greater or equal to $[mT]$ and our procedures test for all potential breakpoints uniformly. A crucial question is how to choose m in practice. The decision might be based on economic arguments and it is possible to support such a choice with a retrospective panel cointegration test as reviewed in Breitung and Pesaran (2008). Our simulations (Figures 1.4 and 1.5) indicate that m should be rather small to have good size properties and it might be good practice to choose an integer multiple of 1/10 such as 0.2 or 0.3.

The following assumption regarding the trend function is typical in fully-modified estimation:

Assumption 1. *There exists a sequence of $p \times p$ scaling matrices $G_{D,T} > 0$, satisfying $||G_{D,T}|| \to 0$ for $T \to \infty$ ($||\cdot||$ any matrix norm), and a p-dimensional vector of functions $D(z)$ with $0 < \int_0^s D(z)D(z)'\mathrm{d}z < \infty$ for $0 \leq s \leq 1$, such that*

$$\lim_{T\to\infty} \sup_{0\leq s\leq 1} \left|\left|T^{1/2}G_{D,T}D_{[sT]} - D(s)\right|\right|_2 = 0 \tag{1.5}$$

for $||\cdot||_2$ the Euclidean norm.

This assumption is essentially the same as in Phillips and Hansen (1990, p. 102) and ensures a well defined limit of the scaled deterministic regressors. In case of a polynomial trend $D_t = [1, t, t^2, \ldots, t^{p-1}]'$ the assumption is satisfied by $G_{D,T} = \mathrm{diag}\left(T^{-1/2}, T^{-1}, \ldots, T^{-p/2}\right)$ and $D(s) = [1, s, s^2, \ldots, s^{p-1}]'$.

Let $\eta_t := [u_t', v_t]'$ be the stacked errors with $u_t = [u_{1,t}, \ldots, u_{N,t}]'$ and $v_t = [v_{1,t}', \ldots, v_{N,t}']'$. Regarding the error process $\{\eta_t\}$ we assume that a functional central limit theorem holds:

Assumption 2.
(a): The stationary process $\{\eta_t\}$ fulfills

$$T^{-1/2}\sum_{t=1}^{[sT]} \eta_t = T^{-1/2}\sum_{t=1}^{[sT]} \begin{bmatrix} u_t \\ v_t \end{bmatrix} \Rightarrow B(s) := BM(\Omega) = \Omega^{1/2}W(s) \tag{1.6}$$

with $W(s) = [W_{u\cdot v}(s)', W_v(s)']'$ an $N(k+1)$-dimensional vector of standard Brownian motions and $0 < \Omega < \infty$, where

$$\Omega = \begin{bmatrix} \Omega_{uu} & \Omega_{uv} \\ \Omega_{vu} & \Omega_{vv} \end{bmatrix} := \sum_{h=-\infty}^{\infty} \mathbb{E}(\eta_0\eta_h'). \tag{1.7}$$

(b): Denoting $S_t^{\eta} := \sum_{j=1}^{t} \eta_j$ it holds that

$$T^{-1} \sum_{t=1}^{[sT]} S_t^{\eta} \eta_t' \Rightarrow \int_0^s B(r) \mathrm{d}B(r)' + \Delta \tag{1.8}$$

with $\Delta := \sum_{h=0}^{\infty} \mathbb{E}(\eta_0 \eta_h')$.

(c): (a) and (b) hold jointly.

$W_{u \cdot v}(s) = [W_{u \cdot v,1}(s), \ldots, W_{u \cdot v,N}(s)]'$ is an N-dimensional vector of standard Brownian motions and $W_v(s) = [W_{v,1}(s)', \ldots, W_{v,N}(s)']'$ consists of N different k-dimensional vectors of standard Brownian motions. We partition $B(s) = [B_u(s)', B_v(s)']'$, where $B_v(s) = [B_{v,1}(s)', \ldots, B_{v,N}(s)']'$ and $B_{v,n}(s)$ is a k-dimensional vector of Gaussian processes for $n = 1, \ldots, N$. The decomposition $B(s) = \Omega^{1/2} W(s)$ holds with

$$\Omega^{1/2} := \begin{pmatrix} (\Omega_{uu} - \Omega_{uv} \Omega_{vv}^{-1} \Omega_{vu})^{1/2} & \Omega_{uv} \Omega_{vv}^{-1/2} \\ \mathbf{0}_{kN \times N} & \Omega_{vv}^{1/2} \end{pmatrix}, \tag{1.9}$$

where $\mathbf{0}_{kN \times N}$ is a kN by N matrix of zeros. The assumption $\Omega_{vv} > 0$ excludes cointegration among the regressors X_t which is typically assumed for FM-OLS estimation.

In order to relate Assumption 2 to the pair of hypotheses (1.3) and (1.4), we call a univariate stochastic process $\{\xi_t\}_{t \in \mathbb{Z}}$ I(0) if it fullfills (potentially after demeaning) a functional central limit theorem, that means, if it holds for $0 \leq s \leq 1$ that $T^{-1/2} \sum_{t=1}^{[sT]} \xi_t \Rightarrow \omega W(s)$, where $W(s)$ denotes a standard Brownian motion and $0 < \omega < \infty$ is the long-run variance $\omega^2 := \sum_{t=-\infty}^{\infty} \mathbb{E}(\xi_0 \xi_t)$ of $\{\xi_t\}_{t \in \mathbb{Z}}$. Therefore, an I(1) process $\{\zeta_t\}_{t \in \mathbb{Z}}$ with $\zeta_t - \zeta_{t-1} = \xi_t$, that is, a summed up I(0) process, fulfills $T^{-1/2} \zeta_{[sT]} \Rightarrow \omega W(s)$ for all $0 \leq s \leq 1$ and ω and $W(s)$ as above.

The monitoring procedures are based on consistent estimators of the parameter vectors θ_n and (co-)variance parameters. Similar to Chu et al. (1996) we assume that structural change occurs only after a break-free calibration period of size $[mT]$ $(0 < m < 1)$ at the beginning of the monitoring in all N cointegrating regressions (c.f. Figure 1.1). We obtain residuals via an estimator $\hat{\theta}_m$ of the fully modified type, which has been studied in detail with regards to panel data structures, based on the calibration period and the corresponding N-dimensional residuals $\hat{u}_{t;m}^{+}$. The monitoring procedure evaluates whether the properly scaled partial sum process of these residuals

$$T^{-1/2}\sum_{t=1}^{[sT]}\hat{u}^{+}_{t;m} \tag{1.10}$$

becomes “too large”. In view of Assumption 2, the partial sum process (1.10) serves as a natural basis for our monitoring procedure. We show that (1.10) converges to a mixture of Gaussian processes with nuisance parameters. The number of nuisance parameters and the precise limiting distribution depend on the set of considered assumptions. Further, the limiting distribution depends on m, the deterministic trend D_t and the number of regressors k as well.

The bases of scalar test statistics are three detectors constructed by different real-valued transformations of the N-dimensional residual process (i.e., mappings from $\mathbb{R}^N$ to $\mathbb{R}$) and we derive their asymptotic behaviour. The detectors[2] are

$$\hat{H}_1^{m,+}(s) := \frac{\left|\left|T^{-1/2}\sum_{t=[mT]+1}^{[sT]}\hat{u}^{+}_{t;m}\right|\right|_2^2}{\left|\left|T^{-1/2}\sum_{t=1}^{[mT]}\hat{u}^{+}_{t;m}\right|\right|_2^2} = \frac{\sum_{n=1}^{N}\left(T^{-1/2}\sum_{t=[mT]+1}^{[sT]}\hat{u}^{+}_{n,t;m}\right)^2}{\sum_{n=1}^{N}\left(T^{-1/2}\sum_{t=1}^{[mT]}\hat{u}^{+}_{n,t;m}\right)^2}, \tag{1.11}$$

$$\begin{aligned}\hat{H}_2^{m,+}(s) &:= \frac{T^{-1}\sum_{i=[mT]+1}^{[sT]}\left|\left|T^{-1/2}\sum_{t=1}^{i}\hat{u}^{+}_{t;m}\right|\right|_2^2}{T^{-1}\sum_{i=1}^{[mT]}\left|\left|T^{-1/2}\sum_{t=1}^{i}\hat{u}^{+}_{t;m}\right|\right|_2^2}\\ &= \frac{\sum_{n=1}^{N}T^{-1}\sum_{i=[mT]+1}^{[sT]}\left(T^{-1/2}\sum_{t=1}^{i}\hat{u}^{+}_{n,t;m}\right)^2}{\sum_{n=1}^{N}T^{-1}\sum_{i=1}^{[mT]}\left(T^{-1/2}\sum_{t=1}^{i}\hat{u}^{+}_{n,t;m}\right)^2}\end{aligned} \tag{1.12}$$

and

$$\begin{aligned}\hat{H}_3^{m,+}(s) &:= \frac{\left|\left|T^{-1}\sum_{i=[mT]+1}^{[sT]}T^{-1/2}\sum_{t=1}^{i}\hat{u}^{+}_{t;m}\right|\right|_2^2}{\left|\left|T^{-1}\sum_{i=1}^{[mT]}T^{-1/2}\sum_{t=1}^{i}\hat{u}^{+}_{t;m}\right|\right|_2^2}\\ &= \frac{\sum_{n=1}^{N}\left(T^{-1}\sum_{i=[mT]+1}^{[sT]}T^{-1/2}\sum_{t=1}^{i}\hat{u}^{+}_{n,t;m}\right)^2}{\sum_{n=1}^{N}\left(T^{-1}\sum_{i=1}^{[mT]}T^{-1/2}\sum_{t=1}^{i}\hat{u}^{+}_{n,t;m}\right)^2},\end{aligned} \tag{1.13}$$

with $\hat{u}^{+}_{t;m} = [\hat{u}^{+}_{1,t;m}, \ldots, \hat{u}^{+}_{N,t;m}]$. Clearly, the numerator in (1.12) turns into the detector used by Wagner and Wied (2017) by setting $n = 1$ and into the Shin (1994) statistic by additionally setting $m = 0$ and $s = 1$.

In the following Sections 1.2.1 – 1.2.3 we derive the limiting distribution $\mathbf{W}_{\mathbf{u}\cdot\mathbf{v}}(s)$ of the

[2] A selection from my master thesis (Theising, 2018).

partial sum process (1.10), such that

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{[sT]}\hat{u}_{t;m}^{+} \Rightarrow \mathbf{W}_{\mathbf{u}\cdot\mathbf{v}}(s), \tag{1.14}$$

where $\mathbf{W}_{\mathbf{u}\cdot\mathbf{v}}(s) = [\mathbf{W}_{\mathbf{u}\cdot\mathbf{v},\mathbf{1}}, \dots, \mathbf{W}_{\mathbf{u}\cdot\mathbf{v},\mathbf{N}}]'$ is a functional of Brownian motions and depends on the set of assumptions specified in the respective sections. All detectors are continuous mappings of the scaled partial sum process (1.10) and we use (1.14) along with the continuous mapping theorem to construct statistical hypothesis tests by the following Lemma 1 which is valid for three sets of assumptions and refers to three different convergence results for (1.14), Lemma 3 in Section 2.1 for homogeneity and cross-sectional independence, Lemma 5 in Section 2.2 for homogeneity and cross-sectional dependence and Lemma 7 in Section 2.3 for heterogeneity and cross-sectional dependence.

Lemma 1. *Let the assumptions of Lemma 3, 5 or 7 be in place, respectively, and let N be fixed. Then it holds that*

$$\hat{H}_1^{m,+}(s) \Rightarrow \frac{\sum_{n=1}^{N}\left(\mathbf{W}_{\mathbf{u}\cdot\mathbf{v},\mathbf{n}}(s) - \mathbf{W}_{\mathbf{u}\cdot\mathbf{v},\mathbf{n}}(m)\right)^2}{\sum_{n=1}^{N}\mathbf{W}_{\mathbf{u}\cdot\mathbf{v},\mathbf{n}}(m)^2} =: \mathcal{H}_1^{m,+}(s), \tag{1.15}$$

$$\hat{H}_2^{m,+}(s) \Rightarrow \frac{\sum_{n=1}^{N}\int_m^s\left(\mathbf{W}_{\mathbf{u}\cdot\mathbf{v},\mathbf{n}}(t)\right)^2 \mathrm{d}t}{\sum_{n=1}^{N}\int_0^m\left(\mathbf{W}_{\mathbf{u}\cdot\mathbf{v},\mathbf{n}}(t)\right)^2 \mathrm{d}t} =: \mathcal{H}_2^{m,+}(s) \tag{1.16}$$

and

$$\hat{H}_3^{m,+}(s) \Rightarrow \frac{\sum_{n=1}^{N}\left(\int_m^s\mathbf{W}_{\mathbf{u}\cdot\mathbf{v},\mathbf{n}}(t)\mathrm{d}t\right)^2}{\sum_{n=1}^{N}\left(\int_0^m\mathbf{W}_{\mathbf{u}\cdot\mathbf{v},\mathbf{n}}(t)\mathrm{d}t\right)^2} =: \mathcal{H}_3^{m,+}(s), \tag{1.17}$$

for $T \to \infty$, where the specific form of $\mathbf{W}_{\mathbf{u}\cdot\mathbf{v},\mathbf{n}}(s)$ depends on which of the three situations is considered. (1.15) - (1.17) are only valid provided the denominators of (1.11) - (1.13) and their respective limits are invertible.

A monitoring procedure rejects the null hypothesis if the weighted detector $\frac{\hat{H}^{m,+}(s)}{g(s)}$ exceeds a critical value for the first time, where $\hat{H}^{m,+}(s)$ is any of the above detectors and $g(s)$ is a weighting function that has to be chosen. This point in time is

$$\tau_m := \min_{s:[mT]+1\leq[sT]\leq T}\left\{\frac{\hat{H}^{m,+}(s)}{g(s)} > c\right\} \tag{1.18}$$

and called detection time. If $\frac{\hat{H}^{m,+}(s)}{g(s)} \leq c$ for all $m \leq s \leq 1$, we set $\tau_m := \infty$. Hence, a finite value of τ_m implies a rejection of the null hypothesis and serves as an immediate estimate of the potential breakpoint. The critical value c and the weighting function $g(s)$ have to be chosen such that under the null hypothesis it holds that

$$\begin{aligned}\lim_{T\to\infty} \mathbb{P}\left(\tau_m < \infty\right) &= \lim_{T\to\infty} \mathbb{P}\left(\min_{s:[mT]+1\le[sT]\le T}\left\{\frac{\hat{H}^{m,+}(s)}{g(s)} > c\right\} < \infty\right)\\ &= \lim_{T\to\infty} \mathbb{P}\left(\sup_{s:[mT]+1\le[sT]\le T}\frac{\hat{H}^{m,+}(s)}{g(s)} > c\right)\\ &= \mathbb{P}\left(\sup_{m\le s\le 1}\frac{\mathcal{H}^{m,+}(s)}{g(s)} > c\right) = \alpha,\end{aligned} \tag{1.19}$$

where α denotes the chosen significance level (c.f. Wagner and Wied, 2017). We only allow for continuous, positive and bounded weighting functions. Clearly, τ_m and c depend on the chosen detector as well as on m, the deterministic trend D_t and the number of regressors k. According to (1.19), the decision rule to reject the null hypothesis if $\tau_m < \infty$ is equivalent to rejecting the null hypothesis if $\sup_{s:[mT]+1\le[sT]\le T}\frac{\hat{H}^{m,+}(s)}{g(s)} > c$.

Using the limits established above and the continuous mapping theorem we derive (see Wagner and Wied, 2017):

Theorem 1. *Let the assumptions of Lemma 1 be in place and assume that $g(s)$ is continuous with $0 < g(s) < \infty$ for $m \le s \le 1$. Then, under the null hypothesis there exist for any $0 < \alpha < 1$ critical values $c = c(\alpha, g, \hat{H}_i^{m,+})$, such that*

$$\lim_{T\to\infty} \mathbb{P}\left(\tau_m(g, c(\alpha, g, \hat{H}_i^{m,+})) < \infty\right) = \alpha, \tag{1.20}$$

for $i = 1, \ldots, 3$.

We calculate the order of the expected value of the three limit processes to motivate our choice of $g(s)$ for intercept and linear trend since optimal weighting functions, for example, in the sense of minimum detection delay, are in general not deducible (see Chu et al., 1996). Hence, we use the monoms matching the respective detector displayed in Table 1.1 for the cases intercept only or linear trend, and arbitrary number of regressors k. In order to obtain critical values $c(\alpha, g, \hat{H}_i^{m,+})$ we need to simulate the limiting distribution $\frac{\mathcal{H}_i^{m,+}(s)}{g(s)}$ by approximating functionals of Brownian motions by the corresponding functions of random walks (see Appendix 1.B for details).

Finally, we introduce some additional notation. With the stacked errors $\eta_{n,t} := [u_{n,t}, v_{n,t}']'$ associated to individual cointegrating regressions we define long-run and one-sided long-

Detector	$D_t = 1$	$D_t = [1,t]'$
$\mathbb{E}(\mathcal{H}_1^{m,+}(s))$	s^2	s^4
$\mathbb{E}(\mathcal{H}_2^{m,+}(s))$	s^3	s^5
$\mathbb{E}(\mathcal{H}_3^{m,+}(s))$	s^4	s^6

Table 1.1.: Order of the expected values of the limiting distributions (1.15) - (1.17) in the case of intercept only ($D_t = 1$) or linear trend ($D_t = [1,t]'$) and no regressors, $\dim X_t = 0$.

run covariances of $\eta_{n,t}$ as

$$\begin{aligned}
\Omega^{m,n} &= \begin{pmatrix} \Omega^{m,n}_{uu} & \Omega^{m,n}_{uv} \\ \Omega^{m,n}_{vu} & \Omega^{m,n}_{vv} \end{pmatrix} := \sum_{h=-\infty}^{\infty} \mathbb{E}(\eta_{m,0}\eta'_{n,h}), \\
\Delta^{m,n} &= \begin{pmatrix} \Delta^{m,n}_{uu} & \Delta^{m,n}_{uv} \\ \Delta^{m,n}_{vu} & \Delta^{m,n}_{vv} \end{pmatrix} := \sum_{h=0}^{\infty} \mathbb{E}(\eta_{m,0}\eta'_{n,h}), \\
\Omega^{m,\cdot}_{ij} &:= [\Omega^{m,1}_{ij}, \dots, \Omega^{m,N}_{ij}]
\end{aligned}$$

and

$$\Delta^{m,\cdot}_{ij} := [\Delta^{m,1}_{ij}, \dots, \Delta^{m,N}_{ij}]$$

for $i, j \in \{u, v\}$ and $m, n = 1, \dots, N$. In what follows we denote consistent estimators of the (one-sided) long-run variance based on the calibration period with a subscript m and "$\wedge$" on top. By $(A)_{n,\cdot}$ we denote the n-th row of a matrix A. $\mathbf{I}_n$ is an $n \times n$ unity matrix and $\mathbf{1}_{n\times m}$ is an $n \times m$ matrix of ones.

1.2.1. Uncorrelated Homogeneous Cointegrating Regressions

We revert to homogeneous cointegrating relationships by imposing the following additional assumption of cross-sectionally identical parameters:

Assumption 3. *$\theta_D = \theta_{D,n}$ and $\theta_X = \theta_{X,n}$ for all $n = 1, \dots, N$.*

Assumption 3 implies

$$y_{n,t} = \begin{cases} D'_t\theta_D + X'_{n,t}\theta_X + u_{n,t}, & t = 1, \dots, [rT], \\ D'_t\theta_{D,1,n} + X'_{n,t}\theta_{X,1,n} + u_{n,t}, & t = [rT]+1, \dots, T, \end{cases} \tag{1.21}$$

and

$$\Delta X_{n,t} = v_{n,t}, \quad t = 1, \ldots, T, \tag{1.22}$$

for $n = 1, \ldots, N$, and simplifies the null hypothesis and alternative with regards to the parameters θ_D and θ_X. Under the null hypothesis no structural change occurs, i.e., $\theta := [\theta_D', \theta_X']' = [\theta_{D,1,n}', \theta_{X,1,n}']' =: \theta_{1,n}$, and under the alternative there is a change in at least one cointegrating regression. Consequently,

$$H_0 : \begin{cases} \theta = \theta_{1,n} \text{ for all } m \leq r < 1, n = 1, \ldots, N, \text{ and} \\ \{u_{n,t}\}_{t=1,\ldots,T} \text{ is I(0) for all } n = 1, \ldots, N \end{cases} \tag{1.23}$$

and

$$H_1 : \begin{cases} \theta \neq \theta_{1,n} \text{ for some } m \leq r < 1, n \in \{1, \ldots, N\} \textbf{ or} \\ \{u_{n,t}\}_{t=1,\ldots,[rT]} \text{ is I(0) and } \{u_{n,t}\}_{t=[rT]+1,\ldots,T} \text{ is I(1)} \\ \text{for some } m \leq r < 1, n \in \{1, \ldots, N\} \end{cases} \tag{1.24}$$

Note that under the alternative of a parameter change the system may turn heterogeneous, i.e., $\theta_{1,i} \neq \theta_{1,j}$ for some $i, j \in \{1, \ldots, N\}$, or stay homogeneous with $\theta_{1,i} = \theta_{1,j}$ for all $i, j \in \{1, \ldots, N\}$.

With regards to the errors we assume a naive i.i.d. setting at first, namely:

Assumption 4. *The stacked error processes $\{\eta_{n,t} := [u_{n,t}, v_{n,t}']'\}_{t=1,\ldots,T}$ are independent and identically distributed for all n.*

Note that by Assumption 4

$$\Omega^{n,n} = \Omega^{\nu,\nu},$$
$$\Delta^{n,n} = \Delta^{\nu,\nu}$$

for all $n, \nu = 1, \ldots, N$ and

$$\Omega^{n,\nu} = \Delta^{n,\nu} = 0$$

for all $n, \nu = 1, \ldots, N, n \neq \nu$. Due to the above implication the decomposition of $B(s)$

collapses to

$$B(s) = \Omega^{1/2} W(s) = \begin{bmatrix} \mathbf{I_N} \otimes \Omega^{1,1}_{uu} & \mathbf{I_N} \otimes \Omega^{1,1}_{vu} \\ \mathbf{I_N} \otimes \Omega^{1,1}_{uv} & \mathbf{I_N} \otimes \Omega^{1,1}_{vv} \end{bmatrix}^{1/2} W(s),$$

a simpler linear transformation of standard Brownian motions where

$$\Omega^{1/2} := \begin{bmatrix} \mathbf{I_N} \otimes \omega_{u \cdot v} & \mathbf{I_N} \otimes \lambda_{uv} \\ \mathbf{I_N} \otimes \mathbf{0_{k\times 1}} & \mathbf{I_N} \otimes \Omega^{1/2}_{vv} \end{bmatrix} \text{ and } \left(\Omega^{1,1}\right)^{1/2} := \begin{bmatrix} \omega_{u\cdot v} & \lambda_{uv} \\ \mathbf{0_{k\times 1}} & (\Omega^{1,1}_{vv})^{1/2} \end{bmatrix} \tag{1.25}$$

with $\omega^2_{u\cdot v} := \Omega^{1,1}_{uu} - \Omega^{1,1}_{uv}(\Omega^{1,1}_{vv})^{-1}\Omega^{1,1}_{vu}$ and $\lambda_{uv} := \Omega^{1,1}_{uv}(\Omega^{1,1}_{vv})^{-1/2}$. Then $\Omega^{1/2}\left(\Omega^{1/2}\right)' = \Omega$ and

$$\begin{bmatrix} \mathbf{I_N} \otimes \omega_{u\cdot v} & \mathbf{I_N} \otimes \lambda_{uv} \\ \mathbf{I_N} \otimes \mathbf{0_{k\times 1}} & \mathbf{I_N} \otimes (\Omega^{1,1}_{vv})^{1/2} \end{bmatrix} \begin{bmatrix} \mathbf{I_N} \otimes \omega_{u\cdot v} & \mathbf{I_N} \otimes \lambda_{uv} \\ \mathbf{I_N} \otimes \mathbf{0_{k\times 1}} & \mathbf{I_N} \otimes (\Omega^{1,1}_{vv})^{1/2} \end{bmatrix}' = \begin{bmatrix} \mathbf{I_N} \otimes \Omega^{1,1}_{uu} & \mathbf{I_N} \otimes \Omega^{1,1}_{vu} \\ \mathbf{I_N} \otimes \Omega^{1,1}_{uv} & \mathbf{I_N} \otimes \Omega^{1,1}_{vv} \end{bmatrix}$$

hold, respectively. Here, the assumption $\Omega^{1,1}_{vv} > 0$ suffices to exclude cointegration among the regressors $X_{n,t}$ for fixed n which is typically assumed for FM-OLS estimation. Then, Assumption 4 implies no cointegration across X_t.

In order to obtain nuisance parameter free asymptotic distributions of the monitoring statistics we use the PFM-OLS estimator of Phillips and Moon (1999, Section 5.2) for systems of homogeneous cointegrating regressions.[3]

Define $Z_{n,t} := [D'_t, X'_{n,t}]'$ and $Z_t := [Z_{1,t}, \ldots, Z_{N,t}]$ and we have

$$y_t = \begin{bmatrix} y_{1,t} \\ \vdots \\ y_{N,t} \end{bmatrix} = \begin{bmatrix} D'_t & X'_{1,t} \\ \vdots & \vdots \\ D'_t & X'_{N,t} \end{bmatrix} \begin{bmatrix} \theta_D \\ \theta_X \end{bmatrix} + \begin{bmatrix} u_{1,t} \\ \vdots \\ u_{N,t} \end{bmatrix} = Z'_t \theta + u_t$$

due to Assumption 3. Since we assume cross-sectional homogeneity and independence, we modify the dependent variables by using

$$y^{+}_{n,t;m} := y_{n,t} - \hat{\Omega}^{1,1}_{uv;m}(\hat{\Omega}^{1,1}_{vv;m})^{-1}\Delta X_{n,t} \tag{1.26}$$

and

$$\hat{\Delta}^{+}_{vu;m} := \hat{\Delta}^{1,1}_{vu;m} - \hat{\Delta}^{1,1}_{vv;m}(\hat{\Omega}^{1,1}_{vv;m})^{-1}\hat{\Omega}^{1,1}_{vu;m}, \tag{1.27}$$

[3] The following discussion on the PFM-OLS estimator was partially covered in my master thesis (Theising, 2018).

where all estimators indicate the arithmetic mean of the respective non-parametric kernel estimators based on individual cointegrating regressions and the pre-break sample $1, \ldots, [mT]$, e.g., $\hat{\Omega}^{1,1}_{vv;m} = N^{-1} \sum_{n=1}^{N} \hat{\Omega}^{1,1}_{vv,n;m}$, where $\hat{\Omega}^{1,1}_{vv,n;m}$ is based solely on cointegrating regression n. Long-run variances are estimated from the stacked error processes $\hat{\eta}_{n,t} := [\hat{u}_{n,t;m}, v'_{n,t}]'$ for $t = 2, \ldots, [mT]$ where $\hat{u}_{n,t;m}$ are the OLS residuals resulting from equation-wise estimation using the calibration period. We assume that long-run variances are estimated consistently, e.g., under the assumptions of Jansson (2002). The PFM-OLS estimator is given by

$$\begin{aligned}\hat{\theta}_{m,\mathrm{PFM}} :=& \left(\sum_{t=1}^{[mT]} \sum_{n=1}^{N} Z_{n,t} Z'_{n,t}\right)^{-1} \left(\sum_{t=1}^{[mT]} \sum_{n=1}^{N} Z_{n,t} y^{+}_{n,t;m} - N[mT] \begin{bmatrix} \mathbf{0_{p\times 1}} \\ \hat{\Delta}^{+}_{vu;m} \end{bmatrix}\right) \\ =& \left(\sum_{t=1}^{[mT]} Z_t Z'_t\right)^{-1} \left(\sum_{t=1}^{[mT]} Z_t y^{+}_{t;m} - N[mT] \begin{bmatrix} \mathbf{0_{p\times 1}} \\ \hat{\Delta}^{+}_{vu;m} \end{bmatrix}\right),\end{aligned} \tag{1.28}$$

where $y^{+}_{t;m} = [y^{+}_{1,t;m}, \ldots, y^{+}_{N,t;m}]'$.

Note, that Phillips and Moon (1999) consider a panel structure with simultaneously $\{T, N\} \to \infty$, while we confine ourselves to the case $T \to \infty$ and N fixed. Another methodological difference is that they work with random linear error processes (VMA(∞)-processes with random coefficients) and show that they fulfill a panel functional central limit theorem (their Lemma 3) under certain assumptions on the random coefficients (Assumption 1 and 2 in their paper, mainly an i.i.d. assumption and moment conditions; for homogeneous panel cointegration they impose non-random coefficients). Therefore, Phillips and Moon (1999) work with "low-level" assumptions on the error structure while here the "high-level" Assumption 2 states that the errors follow a functional central limit theorem, which in fact is a result based on structural assumptions on the errors. Any set of assumptions that implies Assumption 2 is suitable for our purpose.

Concerning the asymptotic properties of the PFM-OLS estimator we derive the following Lemma (c.f. Phillips and Hansen, 1990):

Lemma 2. *Let the data be generated by (1.21) and (1.22) with Assumptions 1 - 4 in* *place. Then*

$$G_T^{-1}\left(\hat{\theta}_{m,PFM}-\theta\right) \Rightarrow \omega_{u\cdot v}\left(\sum_{n=1}^{N}\int_0^m J_n(r)J_n(r)'\mathrm{d}r\right)^{-1} \times \left(\sum_{n=1}^{N}\int_0^m J_n(r)\mathrm{d}W_{u\cdot v,n}(r)\right), \tag{1.29}$$

as $T \to \infty$ *with* $J_n(r) := [D(r)', B_{v,n}(r)']'$, $G_T := \operatorname{diag}(G_{D,T}, G_{X,T})$ *and* $G_{X,T} := T^{-1}\mathbf{I_k}$.

The corresponding N-dimensional residuals are given by $\hat{u}^+_{t;m,\mathrm{PFM}} := y^+_{t;m} - Z_t'\hat{\theta}_{m,\mathrm{PFM}} = u_t - V_t'(\hat{\Omega}^{1,1}_{vv;m})^{-1}\hat{\Omega}^{1,1}_{vu;m} - Z_t'(\hat{\theta}_{m,\mathrm{PFM}} - \theta)$ with $V_t := [v_{1,t}, \ldots, v_{N,t}]$ and we obtain the following limiting distribution for the scaled partial sum process:[4]

Lemma 3. *Let the data be generated by (1.21) and (1.22) with Assumptions 1 - 4 in place. Then it holds under the null hypothesis and for* $0 \leq s \leq 1$

$$T^{-1/2}\sum_{t=1}^{[sT]}\hat{u}^+_{t;m,PFM} \Rightarrow \omega_{u\cdot v}\Bigg\{W_{u\cdot v}(s) - \int_0^s J^W(r)'\mathrm{d}r\left(\sum_{n=1}^{N}\int_0^m J_n^W(r)J_n^W(r)'\mathrm{d}r\right)^{-1} \times \left(\sum_{n=1}^{N}\int_0^m J_n^W(r)\mathrm{d}W_{u\cdot v,n}(r)\right)\Bigg\} =: \omega_{u\cdot v}\widehat{W}_{u\cdot v}(s) \tag{1.30}$$

for $T \to \infty$ *with* $J^W(r) := [J_1^W(r), \ldots, J_N^W(r)]$ *and* $J_n^W(r) := [D(r)', W_{v,n}(r)']'$.

Note that the process $\widehat{W}_{u\cdot v}(s)$ depends on m, the deterministic trend D_t and the number of regressors k as well but we do not reflect this in our notation.

Under Assumptions 3 and 4 self-normalization cancels out the long-run variance in the detector limit. Hence, we get rid of the well-known and unwanted finite sample size distortions induced by long-run variance estimation. A crucial ingredient here is the homogeneity of long-run variances. Assume for this paragraph that $\omega^2_{u\cdot v,n}$ is the conditional long-run variance in cointegrating relation n and that the conditional long-run variances are not homogeneous across cointegrating relations, i.e., $\omega^2_{u\cdot v,1} \neq \omega^2_{u\cdot v,n}$ for some

[4] This result and the subsequent discussion on homogeneous conditional long-run variances are already part of my master thesis (Theising, 2018).

$n \in \{2, \ldots, N\}$. Then, (1.29) changes to

$$G_T^{-1}(\hat{\theta}_{m,\mathrm{PFM}} - \theta) \Rightarrow \left(\sum_{n=1}^{N} \int_0^m J_n^W(r) J_n^W(r)' \mathrm{d}r\right)^{-1} \left(\sum_{n=1}^{N} \omega_{u \cdot v,n} \int_0^m J_n^W(r) \mathrm{d}W_{u \cdot v,n}(r)\right)$$

for $T \to \infty$, and the j-th component of (1.30), $T^{-1/2} \sum_{t=1}^{[sT]} \hat{u}^{+}_{j,t;m,\mathrm{PFM}}$, converges weakly to

$$\begin{aligned} &\omega_{u \cdot v,j} W_{u \cdot v,j}(s) \\ &\quad - \int_0^s J_j^W(r)' \mathrm{d}r \left(\sum_{n=1}^{N} \int_0^m J_n^W(r) J_n^W(r)' \mathrm{d}r\right)^{-1} \left(\sum_{n=1}^{N} \omega_{u \cdot v,n} \int_0^m J_n^W(r) \mathrm{d}W_{u \cdot v,n}(r)\right) \end{aligned}$$

for $T \to \infty$, where the convergence still holds jointly for all $j = 1, \ldots, N$. The nuisance parameters $\omega_{u \cdot v,n}$ cannot be scaled out in the detectors due to their heterogeneity.

Homogeneous parameters are a crucial assumption for (1.30) as in case of heterogeneous parameters (1.29) is no longer valid and $\hat{\theta}_{m,\mathrm{PFM}}$ is only consistent for the average parameter across all equations (c.f. Phillips and Moon, 1999, p. 1080, remark (c), and recall we do not consider random, but fixed parameters).

Based on Lemma 3 the limiting distributions $\sup_{m \leq s \leq 1} \frac{\mathcal{H}_i^{m,+}(s)}{g(s)}$, $i = 1, 2, 3$, depend on different parameters and we obtain critical values for a selection of them, namely $D_t = 1$ or $D_t = [1, t]'$, using the weighting function corresponding to D_t and the respective detector (c.f. Table 1.1), m-values ranging from 0.1 to 0.9 with mesh 0.01, $N = 1, 2, 3, 5, 10, 20, 30$ and $k = 1, \ldots, 4$. We provide further details on simulating the critical values in Appendix 1.B.

1.2.2. Correlated Homogeneous Cointegrating Regressions

As in Section 1.2.1, the data are generated by (1.21) and (1.22) as we consider monitoring homogeneous cointegrating relationships. Regarding the errors we abandon Assumption 4 of independent and identically distributed error vectors η_t but allow for arbitrary dependence among the regressors – except cointegration among the regressors, i.e., $\Omega_{vv}^{n,n} > 0$ and $\Omega_{vv} > 0$ hold. In this case, the modified dependent variable is $y^{+}_{t;m,\mathrm{GLS}} := y_t - \hat{\Omega}_{uv;m} \hat{\Omega}_{vv;m}^{-1} \Delta X_t$ and due to cross-sectional dependence we use the bias correction term

$$\hat{\delta}_m := \sum_{n=1}^{N} \begin{bmatrix} \mathbf{0_{p\times 1}} \\ (\hat{\Delta}^{n,\cdot}_{vu;m})((\hat{\Omega}^{-1}_{u\cdot v;m})_{n,\cdot})' - \hat{\Delta}^{n,\cdot}_{vv;m}((\hat{\Omega}^{-1}_{u\cdot v;m}\hat{\Omega}_{uv;m}\hat{\Omega}^{-1}_{vv;m})_{n,\cdot})' \end{bmatrix}. \tag{1.31}$$

$\hat{\Omega}_{u\cdot v;m}$ is an estimator of $\Omega_{u\cdot v} := \Omega_{uu} - \Omega_{uv}\Omega_{vv}^{-1}\Omega_{vu}$, the long-run covariance of the modified system error $u^{+}_{t;m,\text{GLS}} := u_t - \hat{\Omega}_{uv;m}\hat{\Omega}^{-1}_{vv;m}\Delta X_t$. In order to deal with an arbitrary error structure, we use the pooled feasible GLS estimator

$$\hat{\theta}_{m,\text{PFM-GLS}} := \left(\sum_{t=1}^{[mT]} Z_t\hat{\Omega}^{-1}_{u\cdot v;m}Z_t'\right)^{-1}\left(\sum_{t=1}^{[mT]} Z_t\hat{\Omega}^{-1}_{u\cdot v;m}y^{+}_{t;m,\text{GLS}} - [mT]\hat{\delta}_m\right) \tag{1.32}$$

of the modified system.

Lemma 4. *Let the data be generated by (1.21) and (1.22) with Assumptions 1 - 3 in place. Then*

$$G_T^{-1}(\hat{\theta}_{m,PFM\text{-}GLS} - \theta) \Rightarrow \left(\int_0^m J(r)\Omega^{-1}_{u\cdot v}J(r)'\mathrm{d}r\right)^{-1}\left(\int_0^m J(r)\Omega^{-1/2}_{u\cdot v}\mathrm{d}W_{u\cdot v}(r)\right) \tag{1.33}$$

as $T \to \infty$ with $G_T = \operatorname{diag}(G_{D,T}, G_{X,T})$, $G_{X,T} = T^{-1}\mathbf{I_k}$, $J(r) := [J_1(r), \dots, J_N(r)]$ and $J_n(r) = [D(r)', B(r)'_{v,n}]'$.

Here, the residual vector is $\hat{u}^{+}_{t;m,\text{PFM-GLS}} := y^{+}_{t;m,\text{GLS}} - Z_t'\hat{\theta}_{\text{PFM-GLS}} = u_t - \hat{\Omega}_{uv;m}\hat{\Omega}^{-1}_{vv;m}v_t - Z_t'(\hat{\theta}_{\text{PFM-GLS}} - \theta)$ and the following Lemma holds for the scaled partial sum process:

Lemma 5. *Let the data be generated by (1.21) and (1.22) with Assumptions 1 - 3 in place. Then it holds under the null hypothesis and for* $0 \le s \le 1$

$$\begin{aligned} T^{-1/2}\sum_{t=1}^{[sT]} \hat{u}^{+}_{t;m,PFM\text{-}GLS} \Rightarrow\ & \Omega^{1/2}_{u\cdot v}W_{u\cdot v}(s) - \int_0^s J(r)'\mathrm{d}r\left(\int_0^m J(r)\Omega^{-1}_{u\cdot v}J(r)'\mathrm{d}r\right)^{-1} \\ & \times \left(\int_0^m J(r)\Omega^{-1/2}_{u\cdot v}\mathrm{d}W_{u\cdot v}(r)\right) \end{aligned} \tag{1.34}$$

as $T \to \infty$ with $J(r) = [J_1(r), \dots, J_N(r)]$ and $J_n(r) = [D(r)', B(r)'_{v,n}]'$.

Taking Lemma 5 into account, the limiting distributions $\sup_{m\le s\le 1}\frac{\mathcal{H}_i^{m,+}(s)}{g(s)}$, $i = 1, 2, 3$, do not only depend on the deterministic trend D_t, the weighting function g and the parameters m, N and k but on the long-run covariance structure as well. This renders tabulating simulated critical values infeasible. In order to perform hypotheses tests we need to estimate Ω consistently and replace nuisance parameters in the limiting distribution by consistent estimators. Then, we simulate critical values under the null hypothesis based on independent copies of $W(s)$ that we can easily transform into $B(s)$ to calculate independent copies of $J(s)$ by plugging in covariance estimates performed on the calibration period (see Appendix 1.B).

1.2.3. Seemingly Unrelated Cointegrating Regressions

Suppose now that the cointegrating regressions have individual parameters and the error vectors are not cross-sectionally independent, i.e., we abandon Assumptions 3 and 4 and the data are generated by (1.1) and (1.2). By defining $\mathbf{Z}_t := \operatorname{diag}(Z_{1,t}, \ldots, Z_{N,t})$ and $\theta := [\theta_1', \ldots, \theta_N']'$ we have

$$y_t = \mathbf{Z}_t'\theta + u_t. \tag{1.35}$$

For fully modified estimation we use the GLS modified dependent variable $y^+_{t;m,\text{GLS}}$ and the bias correction term $\hat{\phi}_m := [\hat{\phi}'_{1;m}, \ldots, \hat{\phi}'_{N;m}]'$ with $\hat{\phi}_{n;m} := [\mathbf{0_{p\times 1}}', ((\Delta^{n,\cdot}_{vu;m})^+)']'$ and

$$(\Delta^{n,\cdot}_{vu;m})^+ := (\hat{\Delta}^{n,\cdot}_{vu;m})((\hat{\Omega}^{-1}_{u\cdot v;m})_{n,\cdot})' - \hat{\Delta}^{n,\cdot}_{vv;m}((\hat{\Omega}^{-1}_{u\cdot v;m}\hat{\Omega}_{uv;m}\hat{\Omega}^{-1}_{vv;m})_{n,\cdot})'. \tag{1.36}$$

Moon (1999) discusses three different estimators for this model where the fully modified SUR estimator

$$\hat{\theta}_{\text{FM-SUR}} := \left(\sum_{t=1}^{[mT]} \mathbf{Z}_t \hat{\Omega}^{-1}_{u\cdot v;m} \mathbf{Z}_t'\right)^{-1} \left(\sum_{t=1}^{[mT]} \mathbf{Z}_t \hat{\Omega}^{-1}_{u\cdot v;m} y^+_{t;m,\text{GLS}} - [mT]\hat{\phi}_m\right), \tag{1.37}$$

is the feasible GLS estimator and efficient among those three estimators (c.f. Park and Ogaki, 1991, for additional details on its efficiency).

Lemma 6. *Let the data be generated by (1.1) and (1.2) with Assumptions 1 and 2 in place. Then*

$$\mathbf{G}_T^{-1}(\hat{\theta}_{FM\text{-}SUR} - \theta) \Rightarrow \left(\int_0^m \mathbf{J}(r)\Omega^{-1}_{u\cdot v}\mathbf{J}(r)'\mathrm{d}r\right)^{-1} \left(\int_0^m \mathbf{J}(r)\Omega^{-1/2}_{u\cdot v}\mathrm{d}W_{u\cdot v}(r)\right) \tag{1.38}$$

as $T \to \infty$ *with* $J_n(r) = [D(r)', B_{v,n}(r)']'$, $\mathbf{J}(r) := \operatorname{diag}(J_1(r), \ldots J_N(r))$ *and* $\mathbf{G}_T := \mathbf{I}_N \otimes \operatorname{diag}(G_{D,T}, G_{X,T})$, $G_{X,T} = T^{-1}\mathbf{I_k}$.

Here, the residual vectors are given by $\hat{u}^{+}_{t;m,\text{FM-SUR}} := y^{+}_{t;m,\text{GLS}} - \mathbf{Z}'_t\hat{\theta}_{\text{FM-SUR}} = u_t - \hat{\Omega}_{uv;m}\hat{\Omega}^{-1}_{vv;m}v_t - \mathbf{Z}'_t(\hat{\theta}_{\text{FM-SUR}} - \theta)$ and the scaled partial sum process of the modified residuals has the following probability limit:

Lemma 7. *Let the data be generated by (1.1) and (1.2) with Assumptions 1 and 2 in place. Then it holds under the null hypothesis and for* $0 \leq s \leq 1$

$$\begin{aligned} T^{-1/2}\sum_{t=1}^{[sT]} \hat{u}^{+}_{t;m,FM\text{-}SUR} \Rightarrow \ & \Omega^{1/2}_{u\cdot v}W_{u\cdot v}(s) - \int_0^s \mathbf{J}(r)'\mathrm{d}r \left(\int_0^m \mathbf{J}(r)\Omega^{-1}_{u\cdot v}\mathbf{J}(r)'\mathrm{d}r\right)^{-1} \\ & \times \left(\int_0^m \mathbf{J}(r)\Omega^{-1/2}_{u\cdot v}\mathrm{d}W_{u\cdot v}(r)\right) \end{aligned} \tag{1.39}$$

as $T \to \infty$ *with* $\mathbf{J}(r) = \operatorname{diag}(J_1(r), \ldots J_N(r))$ *and* $J_n(r) = [D(r)', B(r)'_{v,n}]'$.

Note that, $\sum_{i=1}^{[mT]}\sum_{t=1}^{i}\hat{u}^{+}_{t,\text{FM-SUR}} = 0$ if the regression contains an intercept and a linear trend, and $\sum_{t=1}^{[mT]}\hat{u}^{+}_{t,\text{FM-SUR}} = 0$ if the regression contains an intercept. Thus, (1.15) and (1.17) are not valid in the respective cases.

Lemma 7 yields that the limiting distributions of $\sup_{m\leq s\leq 1}\frac{\mathcal{H}^{m,+}_i(s)}{g(s)}$, $i = 1, 2, 3$, depend on the long-run covariance structure as in Section 1.2.2, hence, tabulating simulated critical values is infeasible. Again, we estimate Ω consistently, replace nuisance parameters in the limiting distribution by consistent estimators and simulate critical values under the null hypothesis based on independent copies of $W(s)$ and covariance estimates from the calibration period.

1.3. Finite Sample Performance

We investigate the finite sample properties of the monitoring procedures based on the different detectors and estimators by means of a simulation study. First, we consider the detectors from Section 2.1 for cross-sectional independence and homogenous parameters, then we move to the detectors from Section 2.2 and 2.3. We extend the data generating process used by Vogelsang and Wagner (2014) and Wagner and Wied (2017):

$$\begin{aligned} y_{n,t} &= \mu + \gamma t + x_{n,t,1}\beta_1 + x_{n,t,2}\beta_2 + u_{n,t}, \\ x_{n,t,i} &= x_{n,t-1,i} + v_{n,t,i}, \quad x_{n,0,i} = 0, \quad i = 1, 2, \end{aligned} \tag{1.40}$$

where

$$\begin{aligned} u_{n,t} &= \rho_1 u_{n,t-1} + \varepsilon_{n,t} + \rho_2(e_{n,t,1} + e_{n,t,2}), \quad u_{n,0} = 0, \\ v_{n,t,i} &= e_{n,t,i} + 0.5 e_{n,t-1,i}, \quad i = 1, 2, \end{aligned} \tag{1.41}$$

for $t = 1, \ldots, [mT]$. $\varepsilon_{n,t}$, $e_{n,t,1}$ and $e_{n,t,2}$ are i.i.d. standard normal random variables and independent of each other. We choose parameter values $\mu = 3, \beta_1, \beta_2, \gamma = 1$ and $\rho_1, \rho_2 \in \{0.3, 0.6\}$. The parameter ρ_2 controls the serial correlation in the regression error $u_{n,t}$ and is set to $\rho_1 = 1$ under the alternative of I(1) errors, while the parameter ρ_2 governs regressor endogeneity ($\rho_2 \neq 0$) or exogeneity ($\rho_2 = 0$).

By this simulation study we investigate which of the detectors $\hat{H}_1^{m,+}$, $\hat{H}_2^{m,+}$ and $\hat{H}_3^{m,+}$ is best in the sense of finite sample size control under the null hypothesis as well as power and detection delay under different alternatives. We are interested in how heterogeneous parameters affect the detectors and investigate what happens if we consider alternatives with different regression parameters cross-sectionally by using alternative parameter estimators discussed in Sections 1.2.2 and 1.2.3. An additional important question is how the detectors perform if structural breaks occur only in a fraction of the cointegrating regressions.

We consider different versions (or in some cases modifications) of (1.40) and (1.41) for $t = [mT] + 1, \ldots, T$ to answer the posed questions. In some scenarios we vary the model in the calibration period $t = 1, \ldots, [mT]$ as well. All hypothesis tests are performed on a 5% significance level and we consider combinations of $m \in \{0.10, 0.11, \ldots, 0.89, 0.90\}$ and $N \in \{2, 3, 5, 10, 20, 30\}$. Note that $\hat{H}_{1,\text{PFM}}^{m,+}$ and $\hat{H}_{3,\text{PFM}}^{m,+}$ do not work for $N = 1$ due to the presence of an intercept and a linear trend in D_t (see Lemma 1 and the remark after Lemma 7 which holds for PFM-OLS and $N = 1$ as well).

1.3.1. Null Rejection Probability in Uncorrelated Homogeneous Cointegrating Regressions

In this section, we analyze the behavior of the detectors from Section 2.1 based on PFM-OLS estimation. We additionally assume (1.40) and (1.41) for $t = [mT] + 1, \ldots, T$ to investigate the finite sample performance under the null hypothesis. In particular we examine if the null rejection probability is reasonably close to the chosen significance level of 5%.

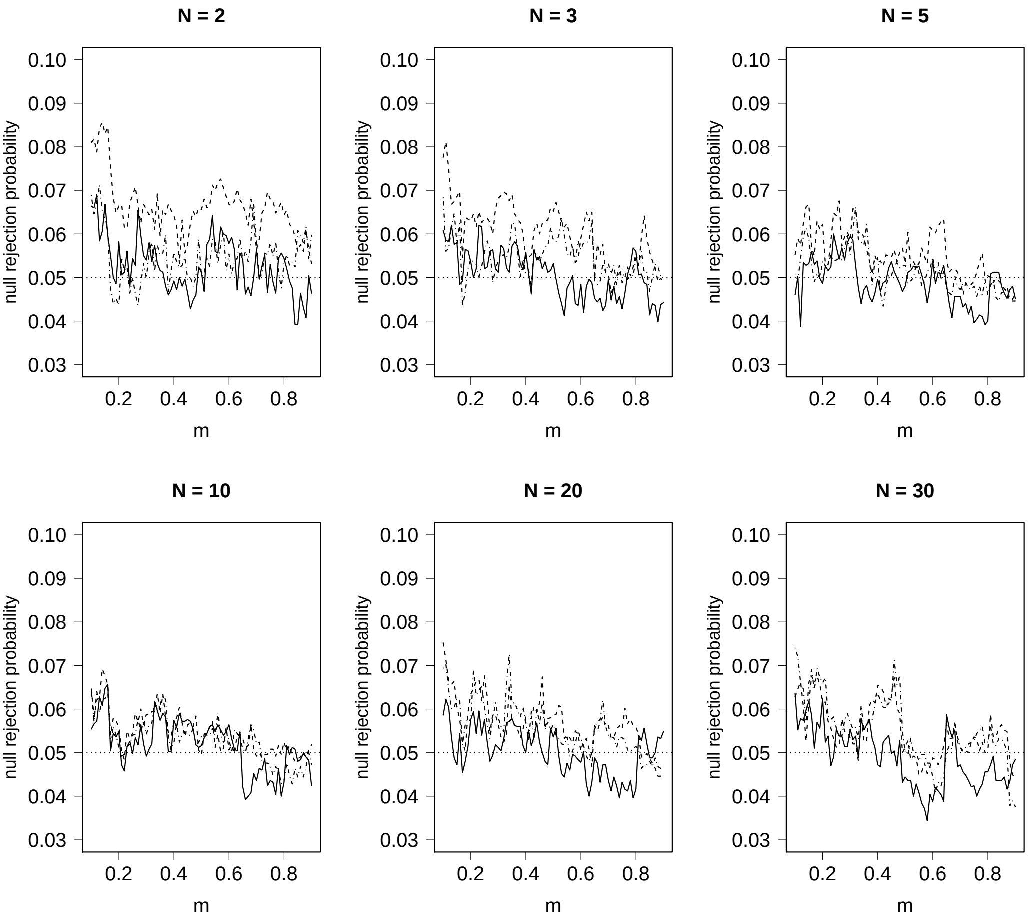

Figure 1.2.: Null rejection probability in uncorrelated homogeneous cointegrating regressions (Section 1.3.1) with $T = 500$, $\rho_1 = \rho_2 = 0.3$ and PFM-OLS estimation. The lines represent $\hat{H}_{1,\mathrm{PFM}}^{m,+}$ (solid), $\hat{H}_{2,\mathrm{PFM}}^{m,+}$ (dashed) and $\hat{H}_{3,\mathrm{PFM}}^{m,+}$ (dotdashed).

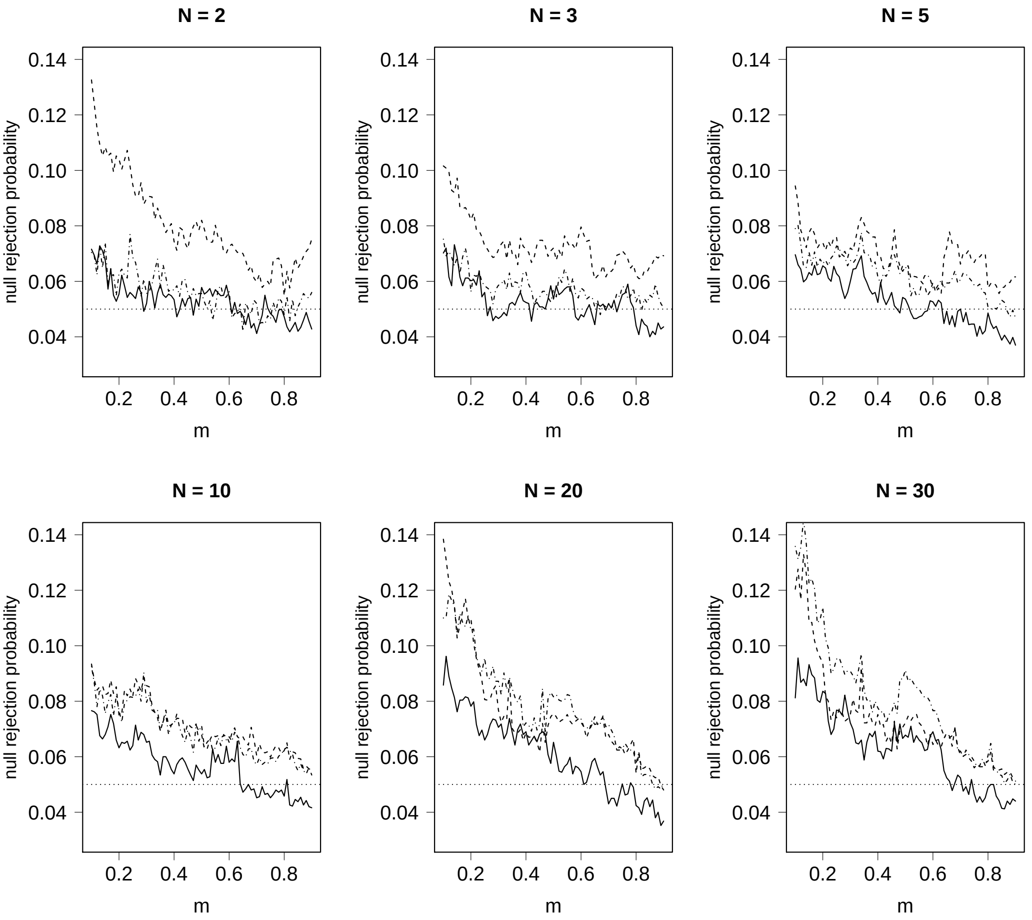

Figure 1.3.: Null rejection probability in uncorrelated homogeneous cointegrating regressions (Section 1.3.1) with $T = 500$, $\rho_1 = \rho_2 = 0.6$ and PFM-OLS estimation. The lines represent $\hat{H}^{m,+}_{1,\mathrm{PFM}}$ (solid), $\hat{H}^{m,+}_{2,\mathrm{PFM}}$ (dashed) and $\hat{H}^{m,+}_{3,\mathrm{PFM}}$ (dotdashed).

In general the null rejection probability is close to the significance level. The detectors suffer from larger long-run variances induced by higher regressor endogeneity and higher error serial correlation. For $N = 2, 3, 5$ the detectors work reasonably well and the null rejection probability decreases in m, size distortions come up for $N = 10$ and get even larger for $N = 20, 30$. All size distortions we observe decrease in m as we use $[mT]$ observations for estimation in the calibration period. $\hat{H}^{m,+}_{1,\mathrm{PFM}}$ seems to perform best in the sense of null rejection probability in this case.

In Figure 1.2 the case $T = 500$ and $\rho_1 = \rho_2 = 0.3$ is shown. The detectors behave similarly well and the null rejection probabilities are close to the significance level ranging between 0.04 and 0.07. In the case of $T = 500$ and $\rho_1 = \rho_2 = 0.6$ (Figure 1.3) $\hat{H}^{m,+}_{1,\mathrm{PFM}}$ and $\hat{H}^{m,+}_{3,\mathrm{PFM}}$ behave similarly and $\hat{H}^{m,+}_{2,\mathrm{PFM}}$ is slightly oversized for $N = 2, 3$. For $N = 5, 10, 20$ $\hat{H}^{m,+}_{2,\mathrm{PFM}}$ and $\hat{H}^{m,+}_{3,\mathrm{PFM}}$ work similarly, slightly above the chosen significance level, only $\hat{H}^{m,+}_{1,\mathrm{PFM}}$ has a lower null rejection probabilty, closer to the significance level. In case of $N = 30$ $\hat{H}^{m,+}_{1,\mathrm{PFM}}$ is closer to the significance level as $\hat{H}^{m,+}_{2,\mathrm{PFM}}$ which is closer than $\hat{H}^{m,+}_{3,\mathrm{PFM}}$. In all cases of N size distortions vanish for larger m.[5]

1.3.2. Null Rejection Probability with Fixed Calibration Period

We looked at finite sample performance by fixing the combined calibration and monitoring period T and compared different sets of parameter values, e.g., the influence of m on the performance of the procedure. This is only helpful in a case of retrospective analysis where we have to specify this value ex post. In the practical case of having a data set and a stream of newly incoming data we cannot specify m a priori independently of T. Merely, we have an assumed to be break free calibration period of a fixed length $[mT]$. Thus, we need to figure out how to specify m and T jointly since there are, in principle, uncountably infinite combinations possible.

In this scenario we simulated under (1.40) and (1.41) for $t = 1, \dots, T$ and applied the detectors from Section 1.2.1. We fixed the value $[mT]$ and simulated time series using pairs (m, T) such that the length of the calibration period is constant displayed in Figures 1.4 and 1.5. The smaller m and consequently the larger T is, the better the performance is in the sense of small size distortion (level 5%). Larger values of T yield better approximations of the test statistics' asymptotic distributions since the procedure

[5] Additional results for $T = 1,000$ in Theising (2018) show a similar pattern with overall smaller size distortions.

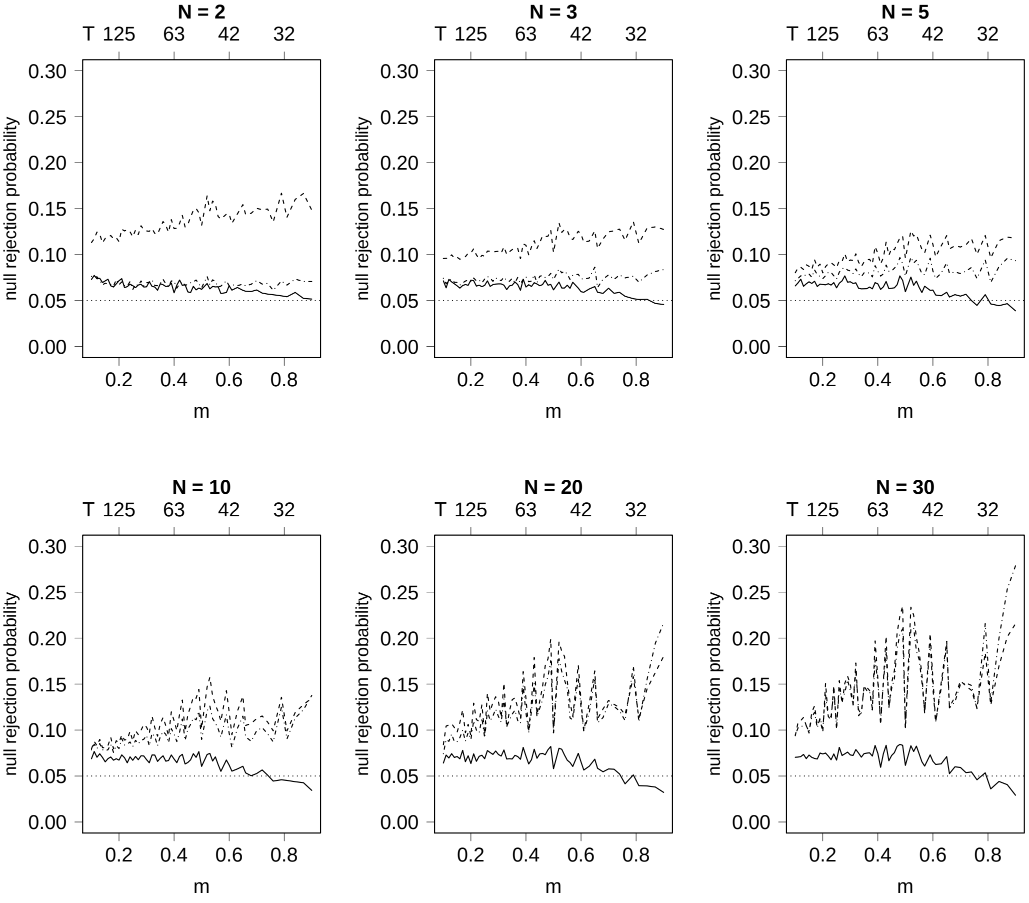

Figure 1.4.: Null rejection probability in uncorrelated homogeneous cointegrating regressions with a fixed calibration period (Section 1.3.2) with $[mT] = 25$, $\rho_1 = \rho_2 = 0.3$ and PFM-OLS estimation. The lines represent $\hat{H}^{m,+}_{1,\mathrm{PFM}}$ (solid), $\hat{H}^{m,+}_{2,\mathrm{PFM}}$ (dashed) and $\hat{H}^{m,+}_{3,\mathrm{PFM}}$ (dotdashed).

is built on large T asymptotics and m is a fixed parameter. That means, we recommend choosing T as large as possible for monitoring newly incoming data.

1.3.3. Null Rejection Probability in Correlated Homogeneous Cointegrating Regressions

We abandon Assumption 4 of independent cointegrating regressions and consider the detectors from Sections 1.2.2 and 1.2.3 based on PFM-GLS and FM-SUR estimation, respectively. We use a data generating process that has a similar covariance and long-run covariance structure as the data in the application section, namely

$$\begin{aligned} y_{n,t} &= \mu + x_{n,t,1}\beta_1 + x_{n,t,2}\beta_2 + u_{n,t}, \\ x_{n,t,i} &= x_{n,t-1,i} + v_{n,t,i}, \quad x_{n,0,i} = 0, \quad i = 1, 2, \end{aligned} \tag{1.42}$$

where

$$\begin{aligned} u_{n,t} &= \rho_1 u_{n,t-1} + (\varepsilon_{n,t} + \rho_2(e_{n,t,1} + e_{n,t,2}))/10, \quad u_{n,0} = 0, \\ v_{n,t,i} &= (e_{n,t,i} + 0.5e_{n,t-1,i} + 0.25e_{n,t-2,i})/10^{3/2}, \quad i = 1, 2, \end{aligned} \tag{1.43}$$

for $t = 1, \dots, [mT]$. $\varepsilon_{n,t}$ is an i.i.d. standard normal random variable independent of $e_t = [e_{1,t,1}, e_{1,t,2}, e_{2,t,1}, \dots e_{N,t,2}]'$. e_t is serially independent and follows a multivariate normal distribution with expected value 0 and covariance matrix $\mathrm{Cov}(e_t) = (1-\tilde{\rho})\mathbf{I_N} + \tilde{\rho}\mathbf{1_{N\times N}}$, where $\mathbf{1_{N\times N}}$ is the $N \times N$ matrix of ones. $\tilde{\rho}$ controls the instantaneous correlation among regressors $v_{n,t}$ and error term $u_{n,t}$ for a single cointegrating regression as well as the instantaneous correlation of regressors and error terms in the cross-section dimension. Further, it holds that $\Omega_{uu} = (1-\rho_1)^{-2}((1+2\rho_2^2(1+\tilde{\rho})-4\tilde{\rho}\rho_2^2)\mathbf{I_N}+4\tilde{\rho}\rho_2^2\mathbf{1_{N\times N}})10^{-2}, \Omega_{vv} = 3.0625((1-\tilde{\rho})\mathbf{I_{kN}} + \tilde{\rho}\mathbf{1_{kN\times kN}})10^{-3}$ and $\Omega_{uv} = 1.75\rho_2(1-\rho_1)^{-1}((1-\tilde{\rho})\mathbf{I_k} \otimes \mathbf{1_{1\times N}} + 2\tilde{\rho}\mathbf{1_{N\times kN}})10^{-5/2}$.

We choose $\mu = 3$ and $\beta_1 = \beta_2 = 1$ again as well as $\tilde{\rho} = 0.9$. The errors are scaled such that they mimic the magnitude and covariance structure of the errors in the application in Section 1.4.

This data generating process violates the assumption of independence across cointegrating relations. By allowing for cross-sectional dependence and heterogeneous (co-)variances the number of additional long-run variance parameters increases from $\frac{1}{2}(k+1)(k+2)$ to $\frac{1}{2}(N+1)k((N+1)k+1)$. A feasible simplification to reduce the number of parameters for large values of N or k is to assume homogeneous long-run variances across $u_{n,t}$ and

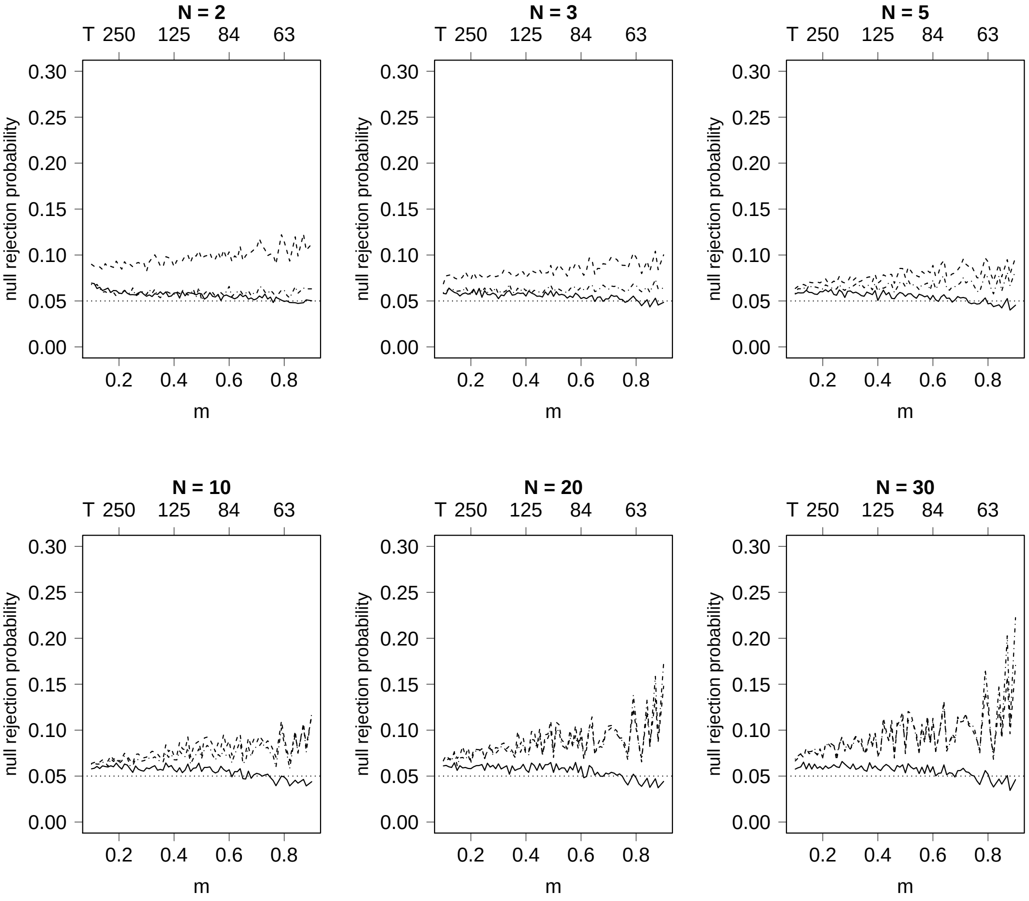

Figure 1.5.: Null rejection probability in uncorrelated homogeneous cointegrating regressions with a fixed calibration period (Section 1.3.2) with $[mT] = 50$, $\rho_1 = \rho_2 = 0.3$ and PFM-OLS estimation. The lines represent $\hat{H}^{m,+}_{1,\text{PFM}}$ (solid), $\hat{H}^{m,+}_{2,\text{PFM}}$ (dashed) and $\hat{H}^{m,+}_{3,\text{PFM}}$ (dotdashed).

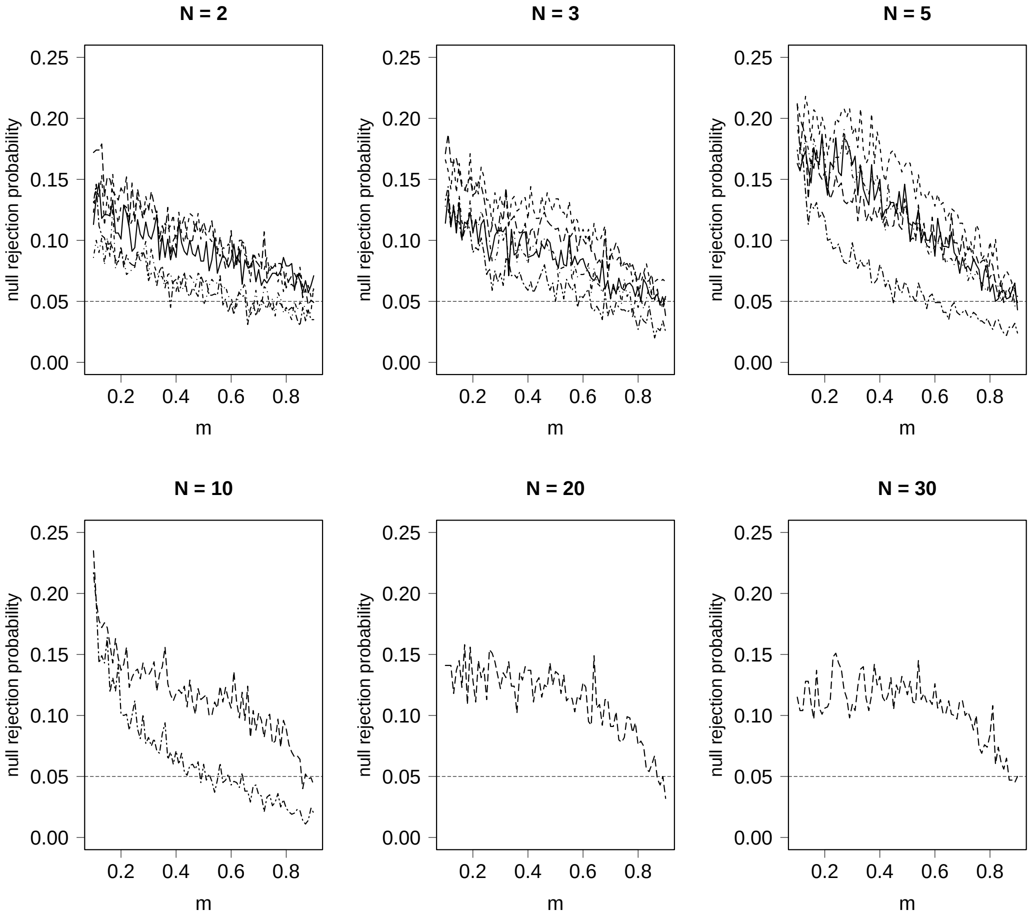

Figure 1.6.: Null rejection probability in correlated homogeneous cointegrating regressions (Section 1.3.3) with $T = 500$, $\rho_1 = \rho_2 = 0.3$, $\tilde{\rho} = 0.9$ and PFM-GLS and FM-SUR estimation. The lines represent $\hat{H}^{m,+}_{1,\text{PFM-GLS}}$ (solid), $\hat{H}^{m,+}_{2,\text{PFM-GLS}}$ (dashed), $\hat{H}^{m,+}_{3,\text{PFM-GLS}}$ (dotdashed), $\hat{H}^{m,+}_{2,\text{FM-SUR}}$ (long-dashed) and $\hat{H}^{m,+}_{3,\text{FM-SUR}}$ (two-dashed).

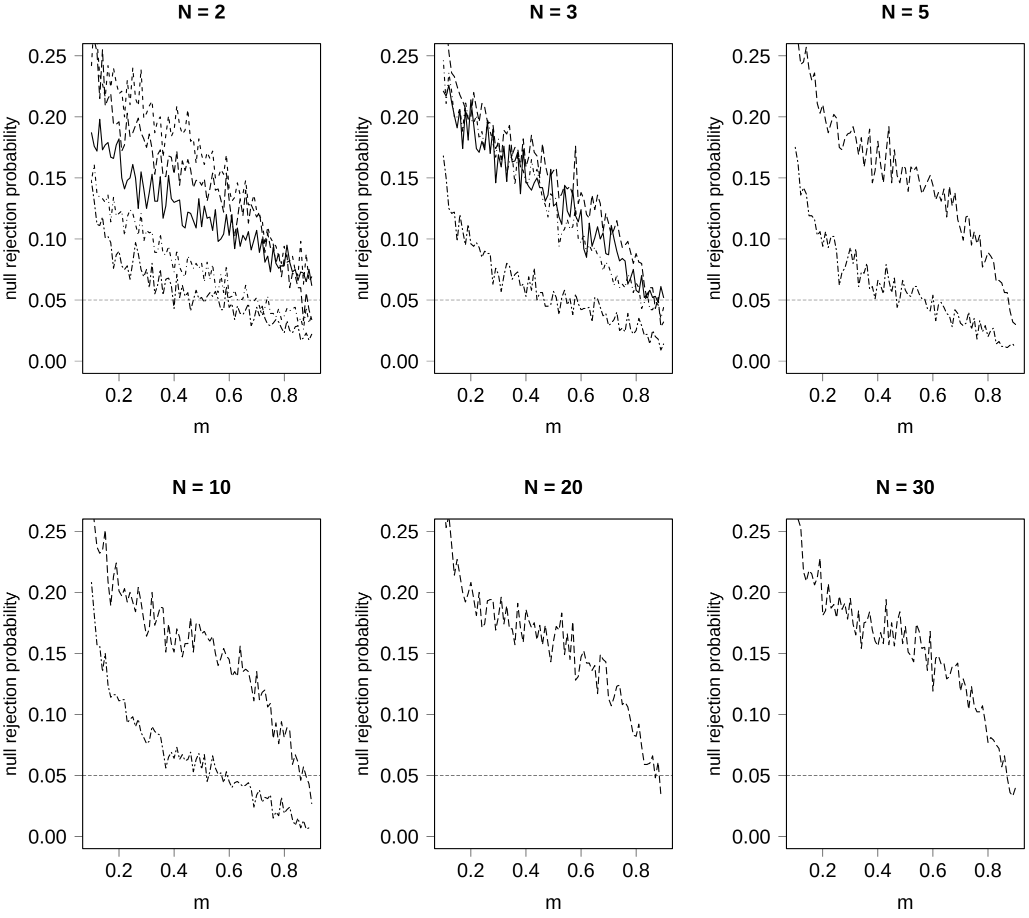

Figure 1.7.: Null rejection probability in correlated homogeneous cointegrating regressions (Section 1.3.3) with $T = 500$, $\rho_1 = \rho_2 = 0.6$, $\tilde{\rho} = 0.9$ and PFM-GLS and FM-SUR estimation. The lines represent $\hat{H}^{m,+}_{1,\text{PFM-GLS}}$ (solid), $\hat{H}^{m,+}_{2,\text{PFM-GLS}}$ (dashed), $\hat{H}^{m,+}_{3,\text{PFM-GLS}}$ (dotdashed), $\hat{H}^{m,+}_{2,\text{FM-SUR}}$ (long-dashed) and $\hat{H}^{m,+}_{3,\text{FM-SUR}}$ (two-dashed).

Table 1.2.: Power in uncorrelated homogeneous cointegrating regressions (Section 1.3.4) with $T = 200$, $\rho_1 = \rho_2 = 0.3$, $N = 5$ and PFM-OLS estimation. The number of breaks is 2 in the first six rows and 4 in the last six rows.

breaks			$\hat{H}^{m,+}_{1,\text{PFM}}$	$\hat{H}^{m,+}_{2,\text{PFM}}$	$\hat{H}^{m,+}_{3,\text{PFM}}$
2	$m = 0.25$	$r = 0.25$	0.14	0.20	0.16
		$r = 0.50$	0.09	0.11	0.08
		$r = 0.75$	0.06	0.07	0.07
	$m = 0.50$	$r = 0.50$	0.42	0.61	0.45
		$r = 0.75$	0.18	0.19	0.10
	$m = 0.75$	$r = 0.75$	0.57	0.64	0.43
4	$m = 0.25$	$r = 0.25$	0.22	0.35	0.25
		$r = 0.50$	0.12	0.16	0.11
		$r = 0.75$	0.08	0.08	0.07
	$m = 0.50$	$r = 0.50$	0.68	0.87	0.72
		$r = 0.75$	0.30	0.33	0.13
	$m = 0.75$	$r = 0.75$	0.81	0.87	0.67

homogeneous long-run variances across $v_{n,t,i}$. Further, assuming $\Omega^{i,j}_{uu} = \Omega^{h,l}_{uu}$, $\Omega^{i,j}_{uv} = \Omega^{h,l}_{uv}$, $\Omega^{i,i}_{uv} = \Omega^{h,h}_{uv}$, $\Omega^{i,j}_{vv} = \Omega^{h,l}_{vv}$ and $\Omega^{i,i}_{vv} = \Omega^{h,h}_{vv}$ for $i \neq j, h \neq l \in \{1, \dots, N\}$ results in $\frac{1}{2}(3k^2 + 5k + 4)$ parameters to be estimated. We realize the estimation of the simplified long-run variance structure for $N > 2$ by pairwisely estimating the long-run variance of all bivariate systems of cointegrating regressions $n_1, n_2 \in \{1, \dots, N\}$ and averaging over all possible pairs.

Figures 1.6 and 1.7 display the null rejection probability for $T = 500$, $\rho_1 = \rho_2 = 0.3, 0.6$, $\tilde{\rho} = 0.9$ and $N \in \{2, 3, 5, 10, 20, 30\}$ of the detectors based on PFM-GLS and FM-SUR estimation. We focus on the detectors which have reasonable empirical sizes. The size distortions are larger than in the case of the PFM-OLS estimator and they depend on which particular estimator is used. The size curves show the interesting pattern that the empirical size declines linearly in m. To some extent, this effect could be expected due to the necessity to estimate nuisance parameters. Still, we believe that the detectors are useful as the size distortions are moderate.

Table 1.3.: Power in uncorrelated homogeneous cointegrating regressions (Section 1.3.4) with $T = 200$, $\rho_1 = \rho_2 = 0.3$, $N = 10$ and PFM-OLS estimation. The number of breaks is 2 in the first six rows and $4, 6, 8, 10$ for each of the following six rows.

breaks			$\hat{H}^{m,+}_{1,\text{PFM}}$	$\hat{H}^{m,+}_{2,\text{PFM}}$	$\hat{H}^{m,+}_{3,\text{PFM}}$
2	$m = 0.25$	$r = 0.25$	0.18	0.24	0.21
		$r = 0.50$	0.10	0.10	0.08
		$r = 0.75$	0.07	0.07	0.06
	$m = 0.50$	$r = 0.50$	0.58	0.61	0.55
		$r = 0.75$	0.23	0.15	0.11
	$m = 0.75$	$r = 0.75$	0.68	0.55	0.39
4	$m = 0.25$	$r = 0.25$	0.35	0.45	0.40
		$r = 0.50$	0.15	0.15	0.12
		$r = 0.75$	0.06	0.07	0.06
	$m = 0.50$	$r = 0.50$	0.86	0.88	0.82
		$r = 0.75$	0.46	0.30	0.16
	$m = 0.75$	$r = 0.75$	0.91	0.82	0.64
6	$m = 0.25$	$r = 0.25$	0.49	0.64	0.56
		$r = 0.50$	0.20	0.22	0.17
		$r = 0.75$	0.08	0.08	0.07
	$m = 0.50$	$r = 0.50$	0.96	0.97	0.93
		$r = 0.75$	0.63	0.44	0.21
	$m = 0.75$	$r = 0.75$	0.98	0.93	0.79
8	$m = 0.25$	$r = 0.25$	0.62	0.76	0.68
		$r = 0.50$	0.28	0.29	0.21
		$r = 0.75$	0.09	0.08	0.08
	$m = 0.50$	$r = 0.50$	0.99	0.99	0.97
		$r = 0.75$	0.77	0.56	0.29
	$m = 0.75$	$r = 0.75$	1.00	0.97	0.88
10	$m = 0.25$	$r = 0.25$	0.72	0.85	0.78
		$r = 0.50$	0.34	0.35	0.26
		$r = 0.75$	0.13	0.10	0.09
	$m = 0.50$	$r = 0.50$	1.00	1.00	0.99
		$r = 0.75$	0.86	0.70	0.36
	$m = 0.75$	$r = 0.75$	1.00	0.99	0.93

1.3.4. Power under Slope Breaks in Uncorrelated Homogeneous Cointegrating Regressions

Now, we turn to power evaluation[6] under slope breaks by which we mean a change in the parameters β_1 or β_2. We consider the detectors from Section 1.2.1 and simulate under (1.40) and (1.41) for $t = 1, \dots, [rT]$ and from $[rT] + 1$ on a subset of the parameters in (1.40) changes. More precisely, there is a break in a different number of the cointegrating relationships and β_1 and β_2 change to $\beta_{1,n} = \beta_{2,n} = 1 - \delta$ in the first half of the breaking cointegrating relationships and change to $\beta_{1,n} = \beta_{2,n} = 1 + \delta$ in the second half with $\delta = 0.05$. Thus, the system is no longer homogeneous after the structural break. Note that we consider $T = 200$ under this alternative as the power is 1 in almost all the cases we study below for $T = 500$ (as under the null hypothesis).

In Table 1.2, we see that $\hat{H}^{m,+}_{2,\mathrm{PFM}}$ has higher power than the other detectors. Keeping in mind that $\hat{H}^{m,+}_{2,\mathrm{PFM}}$ has higher size distortions than $\hat{H}^{m,+}_{1,\mathrm{PFM}}$ and $\hat{H}^{m,+}_{3,\mathrm{PFM}}$ this is no surprise. Between $\hat{H}^{m,+}_{1,\mathrm{PFM}}$ and $\hat{H}^{m,+}_{3,\mathrm{PFM}}$ there is no clear ranking visible regarding power in this scenario. In general, power is increasing in m and higher in the case $m = r$ than in the case $m < r$, i.e., the monitoring works most succesfully when the structural break occurs directly after the end of the calibration period.

Table 1.3 underlines that in most cases $\hat{H}^{m,+}_{2,\mathrm{PFM}}$ has the highest power. The weakness of this detector lies in the case $m = 0.25, r = 0.75$ (or generally in breaks "long" after the calibration period). In this case, $\hat{H}^{m,+}_{1,\mathrm{PFM}}$ has higher power than $\hat{H}^{m,+}_{2,\mathrm{PFM}}$ and $\hat{H}^{m,+}_{3,\mathrm{PFM}}$. All detectors get higher power for higher breakpoint counts where $\hat{H}^{m,+}_{3,\mathrm{PFM}}$ has the worst performance.

1.3.5. Power and Detection Time under Breaks in Uncorrelated Homogeneous Cointegrating Regressions

In this scenario, we consider the detectors from Section 1.2.1 and (1.40) holds for $t = 1, \dots, T$ and (1.41) holds for $t = 1, \dots, [rT]$ where the parameter ρ_1 changes to $\rho_1 = 1$ from $t = [rT] + 1$ on for a subset of the cointegrating relationships. Thus, a fraction of the error processes $\{u_{n,t}\}_{t=[rT]+1,\dots,T}$ are random walks and, therefore, the corresponding cointegrating relationships are no longer valid.

[6] Tables on power of PFM-OLS detectors in 1.3.4 and 1.3.5 are displayed for $T = 200$ in order to allow for a comparison to results for PFM-GLS and FM-SUR in 1.3.6 as opposed to $T = 100$ in Theising and Wied (2023).

Table 1.4.: Power in uncorrelated homogeneous cointegrating regressions (Section 1.3.5) with $T = 200$, $\rho_1 = \rho_2 = 0.3$, $N = 5$ and PFM-OLS estimation. The number of breaks is 1 in the first six rows and $2, 3, 4, 5$ for each of the following six rows.

breaks			$\hat{H}^{m,+}_{1,\text{PFM}}$	$\hat{H}^{m,+}_{2,\text{PFM}}$	$\hat{H}^{m,+}_{3,\text{PFM}}$
1	$m = 0.25$	$r = 0.25$	0.79	0.86	0.73
		$r = 0.50$	0.42	0.46	0.28
		$r = 0.75$	0.12	0.09	0.06
	$m = 0.50$	$r = 0.50$	0.88	0.91	0.79
		$r = 0.75$	0.48	0.46	0.21
	$m = 0.75$	$r = 0.75$	0.80	0.81	0.63
2	$m = 0.25$	$r = 0.25$	0.94	0.97	0.91
		$r = 0.50$	0.65	0.73	0.50
		$r = 0.75$	0.18	0.14	0.08
	$m = 0.50$	$r = 0.50$	0.98	0.99	0.96
		$r = 0.75$	0.73	0.73	0.37
	$m = 0.75$	$r = 0.75$	0.96	0.96	0.84
3	$m = 0.25$	$r = 0.25$	0.99	1.00	0.98
		$r = 0.50$	0.78	0.85	0.62
		$r = 0.75$	0.24	0.16	0.09
	$m = 0.50$	$r = 0.50$	0.99	1.00	0.99
		$r = 0.75$	0.87	0.86	0.49
	$m = 0.75$	$r = 0.75$	0.99	0.99	0.94
4	$m = 0.25$	$r = 0.25$	1.00	1.00	0.99
		$r = 0.50$	0.88	0.93	0.72
		$r = 0.75$	0.29	0.20	0.09
	$m = 0.50$	$r = 0.50$	1.00	1.00	1.00
		$r = 0.75$	0.93	0.92	0.58
	$m = 0.75$	$r = 0.75$	1.00	1.00	0.98
5	$m = 0.25$	$r = 0.25$	1.00	1.00	1.00
		$r = 0.50$	0.93	0.97	0.81
		$r = 0.75$	0.37	0.25	0.10
	$m = 0.50$	$r = 0.50$	1.00	1.00	1.00
		$r = 0.75$	0.96	0.96	0.65
	$m = 0.75$	$r = 0.75$	1.00	1.00	0.99

Table 1.5.: Mean detection delay in uncorrelated homogeneous cointegrating regressions (Section 1.3.5) provided the monitoring procedure detects a break point with $T = 200$, $\rho_1 = \rho_2 = 0.3$, $N = 5$ and PFM-OLS estimation. The number of breaks is 1 in the first six rows and $2, 3, 4, 5$ for each of the following six rows.

breaks			$\hat{H}^{m,+}_{1,\mathrm{PFM}}$	$\hat{H}^{m,+}_{2,\mathrm{PFM}}$	$\hat{H}^{m,+}_{3,\mathrm{PFM}}$
1	$m = 0.25$	$r = 0.25$	39.86	41.97	53.55
		$r = 0.50$	43.46	52.71	57.54
		$r = 0.75$	-7.86	-6.73	-24.19
	$m = 0.50$	$r = 0.50$	31.28	34.36	46.47
		$r = 0.75$	25.94	30.69	31.29
	$m = 0.75$	$r = 0.75$	21.42	25.99	32.90
2	$m = 0.25$	$r = 0.25$	27.67	27.79	42.75
		$r = 0.50$	43.44	49.94	59.96
		$r = 0.75$	6.29	4.23	-16.11
	$m = 0.50$	$r = 0.50$	22.25	24.63	38.74
		$r = 0.75$	24.97	29.42	33.47
	$m = 0.75$	$r = 0.75$	17.28	21.71	29.93
3	$m = 0.25$	$r = 0.25$	21.20	21.45	35.76
		$r = 0.50$	39.82	46.53	58.54
		$r = 0.75$	12.94	14.60	-5.63
	$m = 0.50$	$r = 0.50$	17.06	19.83	33.15
		$r = 0.75$	22.74	28.25	35.24
	$m = 0.75$	$r = 0.75$	14.07	18.86	27.91
4	$m = 0.25$	$r = 0.25$	16.85	17.51	30.09
		$r = 0.50$	38.14	43.68	58.03
		$r = 0.75$	20.78	21.27	1.86
	$m = 0.50$	$r = 0.50$	14.35	17.20	29.52
		$r = 0.75$	21.09	26.01	34.10
	$m = 0.75$	$r = 0.75$	12.14	16.92	26.32
5	$m = 0.25$	$r = 0.25$	13.82	15.15	26.55
		$r = 0.50$	34.57	40.35	56.70
		$r = 0.75$	21.50	23.35	1.95
	$m = 0.50$	$r = 0.50$	12.20	15.45	27.00
		$r = 0.75$	19.51	24.78	33.73
	$m = 0.75$	$r = 0.75$	10.61	15.35	24.61

In Table 1.4 $\hat{H}^{m,+}_{2,\text{PFM}}$ has the highest power in most cases, $\hat{H}^{m,+}_{1,\text{PFM}}$ and $\hat{H}^{m,+}_{3,\text{PFM}}$ behave similarly with some exceptions where $\hat{H}^{m,+}_{3,\text{PFM}}$ has substantially less power. Overall, $\hat{H}^{m,+}_{1,\text{PFM}}$ has power not far off $\hat{H}^{m,+}_{2,\text{PFM}}$ and keeping in mind the size distortions of $\hat{H}^{m,+}_{2,\text{PFM}}$ makes $\hat{H}^{m,+}_{1,\text{PFM}}$ favourable. In general, the power is higher for a higher number of cointegrating relationships with breaks. A weakness lies in the case $m = 0.25$ and $r = 0.75$ where the power is low.

In Table 1.5 we display the mean detection delay conditional on detecting a break point and see that the detection delay is negative in some cases of $m = 0.25$ and $r = 0.75$, suggesting that a lot of false alarms in comparison to correct alarms occur in these cases. The negative delays get closer to 0 or get positive when the number of breaks is greater indicating that the rate of correct alarms gets higher. Given fixed r, the detection delay declines in m with the exception of $m = 0.25, r = 0.75$. Note that the detection delay is bounded by $(m - r)T$ and $(1 - r)T$. The smaller detection delay in the case $m = r = 0.75$ is not contributed to the fact that there are fewer observations left in this case than in the case $m = r = 0.25$ because we see in Table 1.4 that the power is nearly identical in the two cases. Overall, we can say the more breaks occur the smaller the detection delay is. When $m < r$ the detection delay is lower but the power is substantially lower as well. In almost all cases $\hat{H}^{m,+}_{1,\text{PFM}}$ has a smaller detection delay than $\hat{H}^{m,+}_{2,\text{PFM}}$ and $\hat{H}^{m,+}_{3,\text{PFM}}$. $\hat{H}^{m,+}_{2,\text{PFM}}$ detects with a smaller delay than $\hat{H}^{m,+}_{3,\text{PFM}}$ for almost all combinations of m and r with an exception when $m = 0.25$ and $r = 0.75$.

1.3.6. Power and Detection Time under Breaks in Correlated Homogeneous Cointegrating Regressions

We consider the detectors from Sections 1.2.2 and 1.2.3 based on PFM-GLS and FM-SUR estimation, respectively, and revisit the data generating process (1.42) for $t = 1, \dots, T$ and (1.43) for $t = 1, \dots, [rT]$ with cross-sectionally dependent errors. From $[rT] + 1$ onwards for a fraction of the cointegrating relationships the parameter ρ_1 changes to $\rho_1 = 1$ such that a fraction of the error processes $\{u_{n,t}\}_{t=[rT]+1,\dots,T}$ are random walks.

We display our results in Tables 1.6 and 1.7.[7] The PFM-GLS detectors have higher power than the FM-SUR ones, but it should be kept in mind that the empirical size is also higher. Similarly as in the case of Table 1.4, the power is higher for a higher number of cointegrating relationships with breaks and the power is rather low for $m = 0.25$ and

[7] This table is an additional result for this dissertation and not included in Theising and Wied (2023).

Table 1.6.: Power in correlated homogeneous cointegrating regressions (Section 1.3.6) with $T = 200$, $\rho_1 = \rho_2 = 0.3$, $\tilde{\rho} = 0.9$, $N = 5$ and PFM-GLS and FM-SUR estimation. The number of breaks is 1 in the first six rows and $2, 3, 4, 5$ for each of the following six rows.

breaks			$\hat{H}^{m,+}_{1,\text{PFM-GLS}}$	$\hat{H}^{m,+}_{2,\text{PFM-GLS}}$	$\hat{H}^{m,+}_{3,\text{PFM-GLS}}$	$\hat{H}^{m,+}_{2,\text{FM-SUR}}$	$\hat{H}^{m,+}_{3,\text{FM-SUR}}$
1	$m = 0.25$	$r = 0.25$	0.86	0.90	0.81	0.68	0.48
		$r = 0.50$	0.69	0.72	0.55	0.44	0.25
		$r = 0.75$	0.35	0.31	0.21	0.21	0.14
	$m = 0.50$	$r = 0.50$	0.90	0.94	0.84	0.81	0.59
		$r = 0.75$	0.60	0.62	0.34	0.34	0.11
	$m = 0.75$	$r = 0.75$	0.79	0.79	0.62	0.80	0.55
2	$m = 0.25$	$r = 0.25$	0.97	0.98	0.96	0.89	0.70
		$r = 0.50$	0.88	0.90	0.75	0.63	0.32
		$r = 0.75$	0.47	0.36	0.22	0.24	0.15
	$m = 0.50$	$r = 0.50$	0.99	1.00	0.97	0.97	0.85
		$r = 0.75$	0.84	0.84	0.51	0.56	0.20
	$m = 0.75$	$r = 0.75$	0.96	0.96	0.85	0.96	0.79
3	$m = 0.25$	$r = 0.25$	0.99	1.00	0.99	0.97	0.84
		$r = 0.50$	0.95	0.97	0.88	0.74	0.41
		$r = 0.75$	0.56	0.46	0.25	0.22	0.15
	$m = 0.50$	$r = 0.50$	1.00	1.00	0.99	1.00	0.94
		$r = 0.75$	0.94	0.92	0.61	0.70	0.25
	$m = 0.75$	$r = 0.75$	0.99	0.99	0.94	0.99	0.90
4	$m = 0.25$	$r = 0.25$	1.00	1.00	1.00	0.99	0.91
		$r = 0.50$	0.98	0.99	0.94	0.84	0.48
		$r = 0.75$	0.67	0.56	0.26	0.25	0.13
	$m = 0.50$	$r = 0.50$	1.00	1.00	1.00	1.00	0.96
		$r = 0.75$	0.97	0.96	0.69	0.81	0.27
	$m = 0.75$	$r = 0.75$	1.00	1.00	0.97	1.00	0.96
5	$m = 0.25$	$r = 0.25$	1.00	1.00	1.00	1.00	0.94
		$r = 0.50$	0.99	1.00	0.96	0.89	0.53
		$r = 0.75$	0.75	0.63	0.28	0.27	0.16
	$m = 0.50$	$r = 0.50$	1.00	1.00	1.00	1.00	0.99
		$r = 0.75$	0.98	0.99	0.81	0.87	0.34
	$m = 0.75$	$r = 0.75$	1.00	1.00	0.98	1.00	0.98

$r = 0.75$. Mean detection delay in Table 1.7 is again conditional on detecting a break point. Generally, the PFM-GLS detectors have similar detection delay where $\hat{H}^{m,+}_{3,\text{PFM-GLS}}$ is slightly more delayed. Keeping in mind the better empirical behaviour of $\hat{H}^{m,+}_{2,\text{PFM-GLS}}$ under the null hypothesis, this seems to be the favourable detector. All PFM-GLS detectors have smaller detection delay than the FM-SUR counterparts where $\hat{H}^{m,+}_{2,\text{FM-SUR}}$ performs better.

1.4. Application

We consider the cointegrating relation between triplets of logarithmic currency exchange rates. On that account, we calculated exchange rates between Bitcoin and real-world non-cryptocurrencies (USD, EUR, AUD, RUB, etc.) and perform three distinct bivariate analyses meaning that we consider two cointegrating relationships a time. In the analyses, we first consider the detectors from Section 1.2.1. There is statistical evidence that the assumption of cross-sectional independence is not fulfilled, but simulations for robustness using the same data generating process as in Section 1.3.3 indicate that it might still be appropriate to use these detectors (see Appendix 1.C). Under the assumption of cross-sectional dependence, we could use the detectors from Section 1.2.2 and 1.2.3, but would have to estimate additional parameters. Nevertheless, we run the analyses also for the other estimators and observe slightly different results which are also presented.

We use our methods to simultaneously search for instabilities in multiple parities and to our best knowledge there exists no such analysis in the literature, yet. Other authors only consider one currency triplet at a time and therefore just one cointegrating regression. We assume violations of triangular arbitrage parity under normal market conditions to be stationary and a turn to non-stationary deviations or a change in parameters is a sign of mispricing not due to financial frictions – also referred to as financial market dislocation. We find empirical evidence of such mispricing in currency triplets including Bitcoin and use our results in a portfolio trading strategy.

Financial market dislocations are difficult to define and measure, yet arbitrage parities are a less controversial matter (Pasquariello, 2014) and were investigated, for instance, by Yu and Zhang (2017) with a focus on Bitcoin and the relationship between triangular arbitrage parity deviations and cross-country differences in capital controls which can be linked to different demand on foreign currency across countries. Corbet et al. (2018) link Bitcoin prices to fundamentals that seem to drive the price until 2017; after that their

Table 1.7.: Mean detection delay in correlated homogeneous cointegrating regressions (Section 1.3.6) with $T = 200$, $\rho_1 = \rho_2 = 0.3$, $\tilde{\rho} = 0.9$, $N = 5$ and PFM-GLS and FM-SUR estimation. The number of breaks is 1 in the first six rows and $2, 3, 4, 5$ for each of the following six rows.

breaks			$\hat{H}^{m,+}_{1,\text{PFM-GLS}}$	$\hat{H}^{m,+}_{2,\text{PFM-GLS}}$	$\hat{H}^{m,+}_{3,\text{PFM-GLS}}$	$\hat{H}^{m,+}_{2,\text{FM-SUR}}$	$\hat{H}^{m,+}_{3,\text{FM-SUR}}$
1	$m = 0.25$	$r = 0.25$	55.75	57.41	66.98	85.10	95.08
		$r = 0.50$	46.89	53.01	60.06	65.32	69.23
		$r = 0.75$	9.08	7.89	-0.56	11.17	10.54
	$m = 0.50$	$r = 0.50$	38.07	42.38	53.31	55.68	65.51
		$r = 0.75$	26.44	30.39	33.81	36.89	34.45
	$m = 0.75$	$r = 0.75$	24.35	29.40	35.64	30.49	36.62
2	$m = 0.25$	$r = 0.25$	40.47	42.12	55.70	73.67	86.48
		$r = 0.50$	42.47	48.92	60.65	66.62	69.11
		$r = 0.75$	15.56	11.94	1.95	12.13	12.84
	$m = 0.50$	$r = 0.50$	27.22	31.38	43.68	46.18	60.62
		$r = 0.75$	24.42	29.27	35.57	35.47	36.51
	$m = 0.75$	$r = 0.75$	19.40	25.18	32.53	26.13	34.38
3	$m = 0.25$	$r = 0.25$	32.21	33.40	46.08	63.23	80.35
		$r = 0.50$	37.82	43.45	57.10	64.50	69.99
		$r = 0.75$	19.45	18.70	7.84	13.39	12.32
	$m = 0.50$	$r = 0.50$	21.87	25.96	38.12	38.89	55.34
		$r = 0.75$	22.90	27.44	35.00	35.82	37.64
	$m = 0.75$	$r = 0.75$	15.93	22.19	30.47	22.74	31.84
4	$m = 0.25$	$r = 0.25$	26.63	28.28	40.30	53.38	71.99
		$r = 0.50$	35.39	40.42	56.46	61.93	67.99
		$r = 0.75$	20.48	23.33	16.27	17.35	13.98
	$m = 0.50$	$r = 0.50$	19.14	23.70	35.68	35.04	51.99
		$r = 0.75$	20.83	26.03	34.42	34.81	37.13
	$m = 0.75$	$r = 0.75$	14.42	20.20	28.87	20.35	30.30
5	$m = 0.25$	$r = 0.25$	23.56	25.69	36.29	49.41	69.52
		$r = 0.50$	30.97	36.77	52.81	59.81	68.26
		$r = 0.75$	20.59	23.08	12.77	18.17	10.43
	$m = 0.50$	$r = 0.50$	17.09	21.41	32.52	31.91	48.67
		$r = 0.75$	17.70	23.80	34.32	33.71	38.75
	$m = 0.75$	$r = 0.75$	12.81	18.52	26.94	18.70	28.61

model signals bubble-type behaviour. Such bubble-type behaviour can be modeled using the theory of Cretarola and Figà-Talamanca (2021). Cheah and Fry (2015) link Bitcoin prices to fundamentals as well, show that Bitcoin prices are prone to speculative bubbles and find empirical evidence that the fundamental price of Bitcoin is zero; Dong and Dong (2014) conclude that Bitcoin is an immature currency; and Lintilhac and Tourin (2017) use Bitcoin to construct portfolio strategies. Reynolds et al. (2021) investigate the time series properties of Bitcoin and fiat currency logarithmic exchange rates. Their findings suggest that these are unit root processes and they consider univariate cointegrating relationships between triplets of logarithmic currency exchange rates. They present empirical evidence of mispricings in currency triplets including Bitcoin investigating one cointegrating relation at a time and use their result for a currency portfolio strategy.

The law of one price is implied by the assumptions of arbitrage-free markets in modern financial theory meaning prices of related assets are fundamentally linked and should inhibit arbitrage parities. Consider a currency triplet (A-V-B) consisting of three currencies A, B and V (the vehicle currency). Let $S_{A/B,t}$ denote the units of currency A received for one unit of currency B. In the absence of arbitrage, for any triplet of spot exchange rates the triangular arbitrage parity

$$S_{A/B,t} = S_{A/V,t} S_{V/B,t} \quad \Leftrightarrow \quad \ln S_{A/B,t} = \ln S_{A/V,t} + \ln S_{V/B,t} \tag{1.44}$$

holds. In real data we never observe the validity of (1.44). This is suspectedly due to market frictions such as transactions cost. In order to compensate for these frictions we include a stationary error term in (1.44) and assume that deviations from triangular arbitrage parity are stationary transforming (1.44) to

$$\ln S_{A/B,t} = \ln S_{A/V,t} + \ln S_{V/B,t} + u_t, \tag{1.45}$$

where u_t is the stationary error due to market frictions.

Currency triplets sharing more than one currency imply identical regressors and therefore we cannot apply our monitoring procedures: On the one hand, cross-sectional independence would be outruled by construction, on the other hand, a fixed correlation of one would not allow for simplifying the estimator of the long-run variance as described in Section 1.3.3. We consider three examples of two currency triplets a time with Bitcoin (XBT) as vehicle currency V in every triplet. US Dollar (USD) and Euro (EUR) are fixed currencies in each of the two triplets while the third currency varies among Australian Dollar (AUD), Canadian Dollar (CAD), Pound Sterling (GBP), Russian

Ruble (RUB) and Swedish Krona (SEK). This choice is motivated by the fact that USD and EUR can be considered as global leading currencies from large economies and that the chosen currencies are most actively traded in our sample period. The triplets are (USD-XBT-CAD)–(EUR-XBT-GBP), (USD-XBT-SEK)–(GBP-XBT-EUR) and (USD-XBT-AUD)–(RUB-XBT-EUR).

We use daily spot exchange rates among fiat currencies as reported by the Pacific Exchange Rate Service (Bank of Canada, c.f. Antweiler, 2015).The exchange rates are the averages of transaction prices or price quotes from financial institutions between 11:59 a.m. and 12:01 p.m. Eastern time (ET). We use Bitcoin transaction prices between 11:59 a.m. and 12:01 p.m. ET as reported by Bitcoincharts (2017) to calculate noon exchange rates between Bitcoin and fiat currencies.

The chosen triplets leads to three bivariate systems of cointegrating relationships, namely

$$y_t = \begin{bmatrix} \ln S_{\text{USD/CAD},t} \\ \ln S_{\text{EUR/GBP},t} \end{bmatrix} = \begin{bmatrix} 1 & \ln S_{\text{USD/XBT},t} & \ln S_{\text{XBT/CAD},t} \\ 1 & \ln S_{\text{EUR/XBT},t} & \ln S_{\text{XBT/GBP},t} \end{bmatrix} \theta + u_t = X_t' \theta + u_t, \quad (1.46)$$

$$y_t = \begin{bmatrix} \ln S_{\text{USD/SEK},t} \\ \ln S_{\text{GBP/EUR},t} \end{bmatrix} = \begin{bmatrix} 1 & \ln S_{\text{USD/XBT},t} & \ln S_{\text{XBT/SEK},t} \\ 1 & \ln S_{\text{GBP/XBT},t} & \ln S_{\text{XBT/EUR},t} \end{bmatrix} \theta + u_t = X_t' \theta + u_t, \quad (1.47)$$

and

$$y_t = \begin{bmatrix} \ln S_{\text{USD/AUD},t} \\ \ln S_{\text{RUB/EUR},t} \end{bmatrix} = \begin{bmatrix} 1 & \ln S_{\text{USD/XBT},t} & \ln S_{\text{XBT/AUD},t} \\ 1 & \ln S_{\text{RUB/XBT},t} & \ln S_{\text{XBT/EUR},t} \end{bmatrix} \theta + u_t = X_t' \theta + u_t, \quad (1.48)$$

where $\theta = [0, 1, 1]'$ in each of them. The homogeneity of these three systems is a direct consequence of the triangular arbitrage parities.

The sample starts at 1 May 2013 and stretches up to 31 December 2015 due to high Bitcoin trading frequency and thus more reliable Bitcoin prices in this time frame, leading to a small $N = 2$, large $T = 667$ setting. We choose $m = 0.2$, that means calibration ends at 8 November 2013, in order to have a rather small calibration period, compare the discussion after equation (1.4), and assume the cointegrating relation to be break free due to rather stable Bitcoin prices. For a further discussion of this matter the reader is referred to Reynolds et al. (2021). They investigate the time series properties of logarithmic Bitcoin exchange rates and demonstrate that logarithmic exchange rates including Bitcoin behave like I(1) processes. They perform unit root tests (Augmented-Dickey-Fuller and Phillips-Perron) and the KPSS test on logarithmic exchange rates including Bitcoin indicating that they indeed have a unit root. Furthermore, they perform

Table 1.8.: Breakpoint detection dates in the three pairs of currency triplets

	(USD-XBT-CAD) (EUR-XBT-GBP)	(USD-XBT-SEK) (GBP-XBT-EUR)	(USD-XBT-AUD) (RUB-XBT-EUR)
$\hat{H}^{m,+}_{1,\text{PFM}}$	-	-	-
$\hat{H}^{m,+}_{2,\text{PFM}}$	-	09-05-2014	12-02-2015
$\hat{H}^{m,+}_{3,\text{PFM}}$	-	11-07-2014	-
$\hat{H}^{m,+}_{1,\text{PFM-GLS}}$	-	-	-
$\hat{H}^{m,+}_{2,\text{PFM-GLS}}$	-	13-08-2014	-
$\hat{H}^{m,+}_{3,\text{PFM-GLS}}$	-	19-01-2015	-
$\hat{H}^{m,+}_{2,\text{FM-SUR}}$	07-05-2015	16-04-2015	08-05-2015
$\hat{H}^{m,+}_{3,\text{FM-SUR}}$	-	-	-

the same tests on the series of first differences illustrating that these can be assumed to be stationary. For logarithmic exchange rates among fiat currencies I(1) behaviour is well established in the literature.

For monitoring we apply $\hat{H}^{m,+}_{1,\text{PFM}}$, $\hat{H}^{m,+}_{2,\text{PFM}}$ and $\hat{H}^{m,+}_{3,\text{PFM}}$ for the PFM-OLS-case, $\hat{H}^{m,+}_{1,\text{PFM-GLS}}$, $\hat{H}^{m,+}_{2,\text{PFM-GLS}}$ and $\hat{H}^{m,+}_{3,\text{PFM-GLS}}$ for the PFM-GLS-case and $\hat{H}^{m,+}_{2,\text{FM-SUR}}$ and $\hat{H}^{m,+}_{3,\text{FM-SUR}}$ for the FM-SUR-case. In all cases, we use $D_t = 1$ and $g(s)$ according to Table 1.1. We detect structural breaks in all three pairs of currency triplets (c.f. Table 1.8), whereas most breaks can be found for the second triple. Figure 1.8 displays the observed process of all test statistics for all triplets. Important dates for the Bitcoin and financial market during our monitoring and prior to the detected breaks are the shut down of Mt. Gox, a Tokyo-based Bitcoin exchange, in February 2014 (Decker and Wattenhofer, 2014) and the ending of the cap on euro-swiss franc exchange rates by the Swiss National Bank on 15 January 2015.

Given the entanglement of exchange rates the question of independent cointegrating regressions arises naturally. An application of Breusch and Pagan (1980) gives statistical evidence that the assumption of cross-sectional independence might not be reasonable. Similarly as in Section 1.3.3, we have conducted a robustness check investigating the PFM-OLS detectors for dependent cointegrating regressions of the form (1.42) and (1.43). It shows that the detectors work for $N = 2$ even under violation of the independence assumption and behave similarly to the PFM-GLS and FM-SUR detectors for $N = 2$ in this case. In the application examples, the long-run correlation among the first differences of the regressors of different cointegrating regressions $\Delta X_{1,t}$ and $\Delta X_{2,t}$ varies from 0.85 to 0.99 in absolute value and the longrun correlation in (1.43) is 0.9 among these first

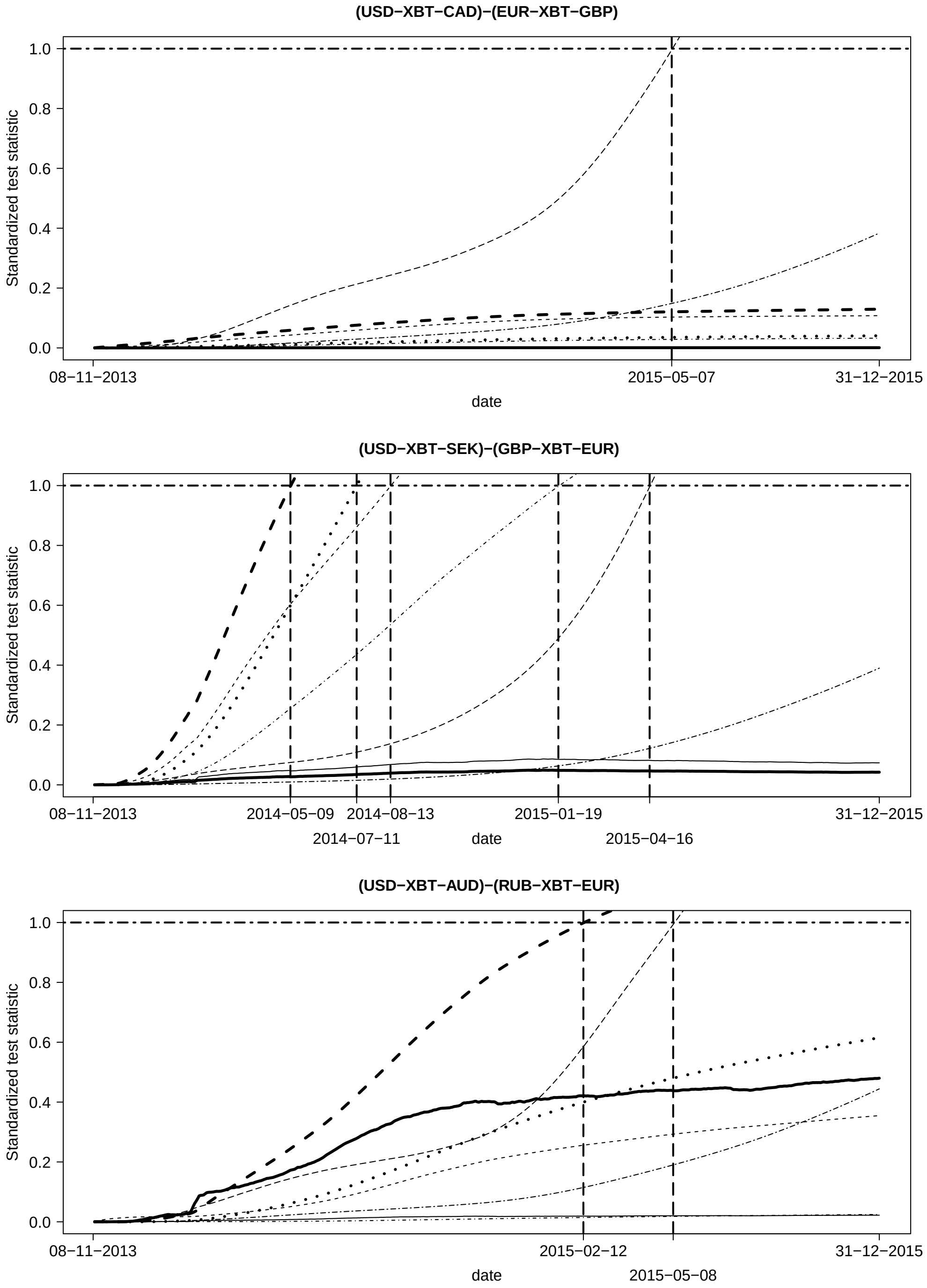

Figure 1.8.: Processes of the test statistics divided by critical values for (1.46) - (1.48). Bold lines represent test statistics based on $\hat{H}^{m,+}_{1,\text{PFM}}$ (solid), $\hat{H}^{m,+}_{2,\text{PFM}}$ (dashed) and $\hat{H}^{m,+}_{3,\text{PFM}}$ (dotdashed), and non-bold lines represent test statistics based on $\hat{H}^{m,+}_{1,\text{PFM-GLS}}$ (solid), $\hat{H}^{m,+}_{2,\text{PFM-GLS}}$ (dashed), $\hat{H}^{m,+}_{3,\text{PFM-GLS}}$ (dotdashed), $\hat{H}^{m,+}_{2,\text{FM-SUR}}$ (long-dashed) and $\hat{H}^{m,+}_{3,\text{FM-SUR}}$ (two-dashed). Dashed vertical lines indicate a detected breakpoint in the system of cointegrating relationships.

Table 1.9.: Returns of buy-and-hold strategy compared to monitoring based portfolio strategy for the three breakpoints

currency triplet	breakdate	benchmark return	excess return
(USD-XBT-CAD) (EUR-XBT-GBP)	07-05-2015	-0.0655	-0.0713
(USD-XBT-SEK) (GBP-XBT-EUR)	09-05-2014	-0.0581	$+0.1211$
	11-07-2014		$+0.2148$
	13-08-2014		$+0.2386$
	19-01-2015		-0.0940
	16-04-2015		-0.1100
(RUB-XBT-EUR) (USD-XBT-AUD)	12-02-2015	-0.1377	-0.0908
	08-05-2015		-0.0451

differences while the variances are of a similar magnitude of 0.0025 in the finite sample case and 0.004 to 0.006 in the application cases. The correlation between the errors $u_{1,t}$ and $u_{2,t}$ in the simulation is 0.2 while the estimated correlation in the applications is between 0.5 and 0.8. The correlations between the first differences of the regressors and the errors $u_{1,t}$ and $u_{2,t}$ vary beween 0.01 and 0.15 in the applications while they are roughly 0.5 in the simulation study. The longrun covariance and correlation matrices of the application examples and details about the econometric test for cross-sectional independence can be found in Appendix 1.C.

We use our results to implement a portfolio trading strategy and compare, first, the three different pairs of triplets and, second, each of the different pairs to a benchmark portfolio using a simple buy-and-hold strategy. Each of the portfolios is equally-weighted among the five currencies included in a pair of currency triplets. USD serves as the domestic currency and we exchange one fifth of the portfolio volume to each of the four foreign currencies present in the respective pair of currency triplets at the start of monitoring on 12 November 2013. We assume that we earn the local risk free rate in each of the currencies which we proxy by the local deposit interest rate given by Euribor (European Money Markets Institute, 2020) as EUR deposit interest rate and LIBOR (Board of Governors of the Federal Reserve System, 2020) as USD deposit interest rate while we obtained AUD, CAD and RUB deposit interest rates from the World Bank (2020) and GBP and SEK deposit interest rates from the Bank of England (2020) and Statistics Sweden (2020), respectively.

The buy-and-hold benchmark portfolios hold the foreign currencies until the end of monitoring on 31 December 2015 and exchange them back to USD. The monitoring based portfolios exchange the foreign currencies back to USD on the detected breakdates and earn the local risk free USD rate until the end of monitoring. We neglect the effects of trading costs. In Table 1.9 we see that in case of the pair (USD-XBT-SEK)–(GBP-XBT-EUR) the monitoring based strategy achieves a substantial excess return compared to the benchmark strategy if we exchange back to USD on any of the first three breaks detected in 2014. Interestingly, trading back to USD on the later detected breaks in 2015 leads to less return compared to the benchmark strategy. As for the pairs (USD-XBT-CAD)–(EUR-XBT-GBP) and (USD-XBT-AUD)–(RUB-XBT-EUR), where only some detectors signal a break, the benchmark strategy generates more return for all three breakdates.

1.5. Summary and Conclusions

We proposed extensions of the monitoring procedures by Wagner and Wied (2017). Again, these extensions are closed-end monitoring procedures designed for a system of cointegrating relationships. Inspired by Chu et al. (1996)parameters are estimated on a break-free calibration period, our procedures are based on the properly scaled partial sum process of residuals and rely on a functional central limit theorem. We use pooled fully modified OLS estimation in order to construct detectors with nuisance parameter free limiting distributions despite error serial correlation and regressor endogeneity in case of homogeneous parameters and independent cointegrating relations. On the one hand, for dependent cointegrating regressions we utilize a pooled fully modified GLS estimator and on the other hand for dependent and heterogeneous cointegrating regressions we employ the fully modified SUR estimator.

In a simulation study it turns out that the detectors show decent behaviour under the null hypothesis with controlled size and have power against two alternatives under different data generating processes. Self-normalization mitigates the impact of long-run variance estimation on the performance of the detectors based on PFM-OLS estimation. Note that, although no estimator of the long-run variance is necessary in these detector, we still need one to perform pooled FM-OLS estimation and obtain residuals. The detectors depend on the assumption of homogeneous parameters and independent cointegrating regressions and under violation of these assumptions PFM-GLS and FM-SUR estimation based detectors show proper behaviour under the null hypothesis as well as under the alternative hypothesis. Note that a higher number of parameters must be estimated for the detectors based on the latter estimators.

As an illustrative application we test for stability in systems of homogeneous cointegrating relationships in triangular arbitrage parities for logarithmic exchange rate triplets including Bitcoin. We use PFM-OLS based detectors as well as PFM-GLS and FM-SUR detectors for monitoring three different examples of bivariate systems of cointegrating relationships in a sample ranging from 1 May 2013 until 31 December 2015 to see if a stochastic version of the triangular arbitrage parity between currency triplets is stable. In one of the cointegrating relationships almost all detectors indicate breakpoints between May 2014 and April 2015. For the other two cointegrating relationships only a small fraction of detectors indicate breaks. Connected events prior to the detected breaks are the closing of Mt. Gox in February 2014 and the ending of the cap on euro-swiss franc exchange rates by the Swiss National Bank in January 2015. We apply these results to construct a portfolio trading strategy using the detected breaks as a sign of currency market instabilities.

Some extensions to this procedure are possible. Insights on the impact of the weighting function on the performance of monitoring procedures are yet to gain. Advantages and disadvantages of the detectors regarding power under specicific alternatives could be analyzed in more detail. The multivariate procedures for monitoring cointegration work best for a small number of cointegration relations and a large number of time periods. Thus, extending the applicability by improving the performance for more cross-sections and fewer time periods is attractive. The self-normalized detectors work better in the multivariate setting and a revisit of the univariate procedure could reveal potential improvements. Finally, methods to deal with non-constant variances, especially in financial applications, are particularly interesting.

1.A. Mathematical Appendix

Proof of Lemma 1.
The lemma follows directly from the continuous mapping theorem and our subsequent results Lemma 3, 5 or 7, respectively. □

Proof of Theorem 1.
For all detectors $\hat{H}_i^{m,+}$, $i = 1, \dots, 3$ the limits $\mathcal{H}_i^{m,+}$ are well defined under the respective assumptions of Sections 1.2.1 – 1.2.3. Analogusly, the limits for $\frac{\hat{H}_i^{m,+}}{g(s)}$ are well defined since $0 < g(s) < \infty$ and $g(s)$ continuous for $0 \leq s \leq 1$. Therefore, critical values for given $g(s)$ can be found for all versions of the detectors (c.f. the proof in Wagner and Wied, 2017). □

Proof of Lemma 2:
The result is stated for $\dim D_t = 0$ in Phillips and Moon (1999, p. 1085) in the first equation after (5.16). Using arguments of Phillips and Hansen (1990) it extends easily to the case of arbitrary deterministic trend D_t satisfying Assumption 1. □

Proof of Lemma 3:[8]
Recall the definition of the N-dimensional PFM-OLS residuals

$$\begin{aligned}\hat{u}_{t;m,\text{PFM}}^{+} =& y_{t;m,\text{PFM}}^{+} - Z_t' \hat{\theta}_{m,\text{PFM}} = y_t - V_t'(\hat{\Omega}_{vv;m}^{1,1})^{-1}\hat{\Omega}_{vu;m}^{1,1} - Z_t'\hat{\theta}_{m,\text{PFM}} \\ =& u_t - V_t'(\hat{\Omega}_{vv;m}^{1,1})^{-1}\hat{\Omega}_{vu;m}^{1,1} - Z_t'(\hat{\theta}_{m,\text{PFM}} - \theta).\end{aligned} \tag{1.49}$$

Consider the decomposition of the PFM-OLS residuals into the above three summands. The limits

$$T^{-1/2}\sum_{t=1}^{[sT]} u_t \Rightarrow \begin{bmatrix} \omega_{u\cdot v} W_{u\cdot v,1}(s) + \Omega_{uv}^{1,1}(\Omega_{vv}^{1,1})^{-1/2} W_{v,1}(s) \\ \vdots \\ \omega_{u\cdot v} W_{u\cdot v,N}(s) + \Omega_{uv}^{1,1}(\Omega_{vv}^{1,1})^{-1/2} W_{v,N}(s) \end{bmatrix}, \tag{1.50}$$

$$T^{-1/2}\sum_{t=1}^{[sT]} V_t'(\hat{\Omega}_{vv;m}^{1,1})^{-1}\hat{\Omega}_{vu;m}^{1,1} \Rightarrow \begin{bmatrix} \Omega_{uv}^{1,1}(\Omega_{vv}^{1,1})^{-1/2} W_{v,1}(s) \\ \vdots \\ \Omega_{uv}^{1,1}(\Omega_{vv}^{1,1})^{-1/2} W_{v,N}(s) \end{bmatrix} \tag{1.51}$$

for $T \to \infty$ are due to Assumption 2 and the consistency of the nonparametric long-run

[8] Revised compared to an incorrect proof in my master thesis Theising (2018).

variance estimators. For the last part, we have

$$
\begin{aligned}
&T^{-1/2}\sum_{t=1}^{[sT]} Z_t'(\hat{\theta}_{m,\mathrm{PFM}}-\theta) = T^{-1/2}\sum_{t=1}^{[sT]}(G_T'Z_t)'G_T^{-1}(\hat{\theta}_{m,\mathrm{PFM}}-\theta)\\
&=\left\{T^{-1}\sum_{t=1}^{[sT]}\begin{bmatrix} T^{1/2}G_{D,T}D_t & \dots & T^{1/2}G_{D,T}D_t \\ T^{1/2}X_{1,t} & \dots & T^{1/2}X_{N,t}\end{bmatrix}'\right\} G_T^{-1}(\hat{\theta}_{m,\mathrm{PFM}}-\theta) \qquad (1.52)\\
&\Rightarrow \omega_{u\cdot v}\int_0^s J^W(r)'\mathrm{d}r\left(\sum_{n=1}^N\int_0^m J_n^W(r)J_n^W(r)'\mathrm{d}r\right)^{-1}\left(\sum_{n=1}^N\int_0^m J_n^W(r)\mathrm{d}W_{u\cdot v,n}(r)\right).
\end{aligned}
$$

for $T\to\infty$. Thus, the asymptotic behaviour of the scaled partial sum process of the PFM-OLS residuals is given by

$$
\begin{aligned}
T^{-1/2}\sum_{t=1}^{[sT]}\hat{u}^+_{t;m,\mathrm{PFM}} \Rightarrow\ & \omega_{u\cdot v}\left\{W_{u\cdot v}(s)-\int_0^s J^W(r)'\mathrm{d}r\left(\sum_{n=1}^N\int_0^m J_n^W(r)J_n^W(r)'\mathrm{d}r\right)^{-1}\right.\\
&\left.\times\left(\sum_{n=1}^N\int_0^m J_n^W(r)\mathrm{d}W_{u\cdot v,n}(r)\right)\right\} = \omega_{u\cdot v}\widehat{W}_{u\cdot v}(s)
\end{aligned}
\qquad (1.53)
$$

as $T\to\infty$. □

Proof of Lemma 4:

The proof is similar to the proof of Lemma 2 with an additional transformation typical for generalized least squares estimators, here by the long-run covariance of the modified system error $u^+_{t;m,\mathrm{GLS}}$. We consider the limit of $G_T^{-1}(\hat{\theta}_{m,\text{PFM-GLS}}-\theta)$ as $T\to\infty$. Recall $y^+_{t;m,\mathrm{GLS}} := y_t-\hat{\Omega}_{uv;m}\hat{\Omega}^{-1}_{vv;m}v_t$ and $y_t = Z_t'\theta+u_t$, then we have

$$
\begin{aligned}
\hat{\theta}_{m,\text{PFM-GLS}} &= \left(\sum_{t=1}^{[mT]} Z_t\hat{\Omega}^{-1}_{u\cdot v;m}Z_t'\right)^{-1}\left(\sum_{t=1}^{[mT]} Z_t\hat{\Omega}^{-1}_{u\cdot v;m}y^+_{t;m,\mathrm{GLS}}-[mT]\hat{\delta}_m\right)\\
&=\left(\sum_{t=1}^{[mT]} Z_t\hat{\Omega}^{-1}_{u\cdot v;m}Z_t'\right)^{-1}\left(\sum_{t=1}^{[mT]} Z_t\hat{\Omega}^{-1}_{u\cdot v;m}(Z_t'\theta+u_t-\hat{\Omega}_{uv;m}\hat{\Omega}^{-1}_{vv;m}v_t)-[mT]\hat{\delta}_m\right) \qquad (1.54)\\
&=\theta+\left(\sum_{t=1}^{[mT]} Z_t\hat{\Omega}^{-1}_{u\cdot v;m}Z_t'\right)^{-1}\left(\sum_{t=1}^{[mT]} Z_t\hat{\Omega}^{-1}_{u\cdot v;m}(u_t-\hat{\Omega}_{uv;m}\hat{\Omega}^{-1}_{vv;m}v_t)-[mT]\hat{\delta}_m\right).
\end{aligned}
$$

The limiting result

$$
\sum_{t=1}^{[mT]} G_TZ_t\hat{\Omega}^{-1}_{u\cdot v}(G_TZ_t)' \Rightarrow \int_0^m J(r)\Omega^{-1}_{u\cdot v}J(r)'\mathrm{d}r \qquad (1.55)
$$

for $T \to \infty$ is analogously obtained as Lemma 3.1 (b) in Phillips and Durlauf (1986) combined with arguments from Phillips and Hansen (1990) for the deterministic part (or c.f. Moon, 1999).

Define $\Sigma := \mathbb{E}(\eta_0\eta_0')$ and $\Delta_1 := \sum_{h=1}^{\infty}\mathbb{E}(\eta_0\eta_h')$ with the same partitions and notations as for Ω and Δ (see Section 1.2) for the parts that involve bias corrections. In order to handle the GLS transformation in $\sum_{t=1}^{[mT]} G_T Z_t \hat{\Omega}_{u\cdot v;m}^{-1} u_t$ we consider one entry of the vector $\hat{\Omega}_{u\cdot v;m}^{-1} u_t$ at a time. Hence, we examine the n-th row of $\hat{\Omega}_{u\cdot v;m}^{-1}$ and the n-th column of the matrix $G_T Z_t$ which we split into its deterministic and random part. Then, we have

$$\begin{aligned}
&\sum_{t=1}^{[mT]} T^{-1} X_{n,t} \left(\hat{\Omega}_{u\cdot v;m}^{-1}\right)_{n,\cdot} u_t \\
&= \sum_{t=1}^{[mT]} T^{-1/2} X_{n,t-1} T^{-1/2} u_t' \left(\left(\hat{\Omega}_{u\cdot v;m}^{-1}\right)_{n,\cdot}\right)' + T^{-1} \sum_{t=1}^{[mT]} v_{n,t} u_t' \left(\left(\hat{\Omega}_{u\cdot v;m}^{-1}\right)_{n,\cdot}\right)' \\
&\Rightarrow \int_0^m B_{v,n}(r) \mathrm{d}B_u(r)' \left(\left(\hat{\Omega}_{u\cdot v;m}^{-1}\right)_{n,\cdot}\right)' + m\Delta_{1;vu}^{n,\cdot} \left(\left(\hat{\Omega}_{u\cdot v;m}^{-1}\right)_{n,\cdot}\right)' + m\Sigma_{vu}^{n,\cdot} \left(\left(\hat{\Omega}_{u\cdot v;m}^{-1}\right)_{n,\cdot}\right)' \\
&= \int_0^m B_{v,n}(r) \left(\hat{\Omega}_{u\cdot v;m}^{-1}\right)_{n,\cdot} \mathrm{d}B_u(r) + m\Delta_{vu}^{n,\cdot} \left(\left(\hat{\Omega}_{u\cdot v;m}^{-1}\right)_{n,\cdot}\right)'
\end{aligned} \tag{1.56}$$

for $T \to \infty$ where we use $\Delta = \Sigma + \Delta_1$. The convergence result is part of the proof of Theorem 3.1 (e) in Phillips and Durlauf (1986). Further, we obtain

$$\sum_{t=1}^{[mT]} T^{1/2} G_T D_t T^{-1/2} u_t' \left(\left(\hat{\Omega}_{u\cdot v;m}^{-1}\right)_{n,\cdot}\right)' \Rightarrow \int_0^m D(r) \left(\hat{\Omega}_{u\cdot v}^{-1}\right)_{n,\cdot} \mathrm{d}B_u(r) \tag{1.57}$$

as $T \to \infty$. Summing these intermediate results up is equivalent to the initial matrix vector product and we have

$$\begin{aligned}
&\sum_{t=1}^{[mT]} G_T Z_t \hat{\Omega}_{u\cdot v;m}^{-1} u_t \\
&\Rightarrow \int_0^m \sum_{n=1}^{N} J_n(r) (\Omega_{u\cdot v}^{-1})_{n,\cdot} \mathrm{d}B_u(r) + m \sum_{n=1}^{N} \begin{bmatrix} \mathbf{0}_{p\times 1} \\ \Delta_{vu}^{n,\cdot} \left((\Omega_{u\cdot v}^{-1})_{n,\cdot}\right)' \end{bmatrix} \\
&= \int_0^m J(r) \Omega_{u\cdot v}^{-1/2} \mathrm{d}W_{u\cdot v}(r) + \int_0^m J(r) \Omega_{u\cdot v}^{-1} \Omega_{uv} \Omega_{vv}^{-1/2} \mathrm{d}W_v(r) + m \sum_{n=1}^{N} \begin{bmatrix} \mathbf{0}_{p\times 1} \\ \Delta_{vu}^{n,\cdot} \left((\Omega_{u\cdot v}^{-1})_{n,\cdot}\right)' \end{bmatrix}
\end{aligned} \tag{1.58}$$

for $T \to \infty$ where we use $\sum_{n=1}^{N} J_n(r)(\Omega_{u\cdot v}^{-1})_{n,\cdot} = J(r)\Omega_{u\cdot v}^{-1}$ and $B_u(r) = \Omega_{u\cdot v}^{1/2} W_{u\cdot v}(r) + \Omega_{uv}\Omega_{vv}^{-1/2} W_v(r)$. By similar matrix manipulation regarding $\hat{\Omega}_{u\cdot v;m}^{-1}\hat{\Omega}_{uv;m}\hat{\Omega}_{vv;m}^{-1}v_t$ we show

$$\begin{aligned}
&\sum_{t=1}^{[mT]} G_T Z_t \hat{\Omega}_{u\cdot v;m}^{-1}\hat{\Omega}_{uv;m}\hat{\Omega}_{vv;m}^{-1}v_t \\
&\Rightarrow \int_0^m J(r)\Omega_{u\cdot v}^{-1}\Omega_{uv}\Omega_{vv}^{-1}\mathrm{d}B_v(r) + m\sum_{n=1}^{N}\begin{bmatrix} \mathbf{0}_{p\times 1} \\ \Delta_{vv}^{n,\cdot}\left((\Omega_{u\cdot v}^{-1}\Omega_{uv}\Omega_{vv}^{-1})_{n,\cdot}\right)' \end{bmatrix} \\
&= \int_0^m J(r)\Omega_{u\cdot v}^{-1}\Omega_{uv}\Omega_{vv}^{-1/2}\mathrm{d}W_v(r) + m\sum_{n=1}^{N}\begin{bmatrix} \mathbf{0}_{p\times 1} \\ \Delta_{vv}^{n,\cdot}\left((\Omega_{u\cdot v}^{-1}\Omega_{uv}\Omega_{vv}^{-1})_{n,\cdot}\right)' \end{bmatrix}
\end{aligned} \tag{1.59}$$

as $T \to \infty$ where we, again, use Theorem 3.1 (e) of Phillips and Durlauf (1986) and $B_v(r) = \Omega_{vv}^{1/2}W_v(r)$. Lastly, we have

$$\begin{aligned}
G_T[mT]\hat{\delta}_m &= \frac{[mT]}{T}\sum_{n=1}^{N}\begin{bmatrix} \mathbf{0_{p\times 1}} \\ (\hat{\Delta}_{vu;m}^{n,\cdot})((\hat{\Omega}_{u\cdot v;m}^{-1})_{n,\cdot})' - \hat{\Delta}_{vv;m}^{n,\cdot}((\hat{\Omega}_{u\cdot v;m}^{-1}\hat{\Omega}_{uv;m}\hat{\Omega}_{vv;m}^{-1})_{n,\cdot})' \end{bmatrix} \\
&\Rightarrow m\sum_{n=1}^{N}\begin{bmatrix} \mathbf{0_{p\times 1}} \\ \Delta_{vu}^{n,\cdot}\left((\Omega_{u\cdot v}^{-1})_{n,\cdot}\right)' - \Delta_{vv}^{n,\cdot}\left((\Omega_{u\cdot v}^{-1}\Omega_{uv}\Omega_{vv}^{-1})_{n,\cdot}\right)' \end{bmatrix}
\end{aligned} \tag{1.60}$$

for $T \to \infty$. Combining all limits completes the proof

$$G_T^{-1}(\hat{\theta}_{m,\text{PFM-GLS}} - \theta) \Rightarrow \left(\int_0^m J(r)\Omega_{u\cdot v}^{-1}J(r)'\mathrm{d}r\right)^{-1}\left(\int_0^m J(r)\Omega_{u\cdot v}^{-1/2}\mathrm{d}W_{u\cdot v}(r)\right) \tag{1.61}$$

for $T \to \infty$. □

Proof of Lemma 5:

Besides the different estimation technique using generalized least squares transformation the proof is similar to the proof of Lemma 3. The decomposition of the N-dimensional residual vector in this case is $\hat{u}_{t;m,\text{PFM-GLS}}^{+} = u_t - \hat{\Omega}_{uv;m}\hat{\Omega}_{vv;m}^{-1}v_t - Z_t'(\hat{\theta}_{m,\text{PFM-GLS}} - \theta)$. Then,

$$T^{-1/2}\sum_{t=1}^{[sT]} u_t \Rightarrow \Omega_{u\cdot v}^{1/2}W_{u\cdot v}(s) + \Omega_{uv}\Omega_{vv}^{-1/2}W_v(s) \tag{1.62}$$

and

$$T^{-1/2}\sum_{t=1}^{[sT]} \hat{\Omega}_{uv;m}\hat{\Omega}_{vv;m}^{-1}v_t \Rightarrow \Omega_{uv}\Omega_{vv}^{-1/2}W_v(s) \tag{1.63}$$

hold as $T \to \infty$. For the last part we multiply by $G_T G_T^{-1}$ again and obtain the result

$$\begin{aligned} T^{-1/2} \sum_{t=1}^{[sT]} Z_t'(\hat{\theta}_{m,\text{PFM-GLS}} - \theta) &= T^{-1/2} \sum_{t=1}^{[sT]} (G_T Z_t)' G_T^{-1} (\hat{\theta}_{m,\text{PFM-GLS}} - \theta) \\ &\Rightarrow \int_0^s J(r)' \mathrm{d}r \left(\int_0^m J(r) \Omega_{u \cdot v}^{-1} J(r)' \mathrm{d}r \right)^{-1} \left(\int_0^m J(r) \Omega_{u \cdot v}^{-1/2} \mathrm{d}W_{u \cdot v}(r) \right) \end{aligned} \tag{1.64}$$

as $T \to \infty$ by using Lemma 4. Adding all three parts completes the proof. □

Proof of Lemma 6:
This result is stated for $\dim D_t = 0$ in Moon (1999) and is easily extended to the case of arbitrary deterministic trend fulfilling Assumption 1 by arguments of Phillips and Hansen (1990). □

Proof of Lemma 7:
The proof is similar to the proof of Lemma 5. □

1.B. Simulating Critical Values

In order to obtain asymptotically size controlled monitoring procedures we need critical values. Therefore, we simulate quantiles of $\sup_{m \le s \le 1} \{ \frac{\mathcal{H}(s)}{g(s)} \}$, where $\mathcal{H}(s)$ is any of the limiting distributions of the detectors in Lemma 1 and $g(s)$ is the corresponding weighting function (see Table 1.1). The limiting distributions of the detectors are functionals of the limit processes $\mathbf{W}_{\mathbf{u} \cdot \mathbf{v}}(s)$ of the scaled partial sum process $\frac{1}{\sqrt{T}} \sum_{t=1}^{[sT]} \hat{u}_{t;m}^{+}$. Consider the case of PFM-GLS estimation in Section 1.2.2. Then, the limit process is

$$\Omega_{u \cdot v}^{1/2} W_{u \cdot v}(s) - \int_0^s J(r)' \mathrm{d}r \left(\int_0^m J(r) \Omega_{u \cdot v}^{-1} J(r)' \mathrm{d}r \right)^{-1} \left(\int_0^m J(r) \Omega_{u \cdot v}^{-1/2} \mathrm{d}W_{u \cdot v}(r) \right),$$

c.f. (1.34) and Lemma 5. This process is a functional of vectors of independent standard Brownian motions $W_{u \cdot v}(s)$ and $W_v(s)$ independent of each other (recall $J(s) = [J_1(s), \ldots, J_N(s)]$, $J_n(s) = [D(s)', B(s)_{v,n}']'$ and $B_v(s) = \Omega_{vv}^{1/2} W_v(s)$). We approximate functionals of standard Brownian motions using the corresponding functions of random walks of length 1,000 generated from i.i.d. standard normal random variables. We justify this by the functional central limit theorem $\mathcal{W}_{v,n}(s) := T^{-1/2} \sum_{j=1}^{[sT]} \mathcal{X}_{j,v,n} \Rightarrow W_{v,n}(s)$ for $T \to \infty$, where $\mathcal{X}_{j,v,n}$ are k-dimensional i.i.d. random vectors with independent standard normal entries and $W_{v,n}(s)$ is a k-dimensional standard Brownian motion for $n = 1, \ldots, N$. We argue that $T = 1,000$ should be large enough in order for $\mathcal{W}_{v,n}(s)$ to behave approximately like a k-dimensional vector of independent standard Brownian motions.

Turning to the three integrals, consider the components of the first integral $\int_0^s J_n(r)'\mathrm{d}r = \int_0^s [D(r)', B_{v,n}(r)']'\mathrm{d}r$. Since we know the deterministic function $D(s)$ beforehand we can calculate the integral $\int_0^s D(r)'\mathrm{d}r$ analytically but $\int_0^s B_{v,n}(r)'\mathrm{d}r$ needs to be approximated numerically as well as $\int_0^m J(r)\Omega_{u\cdot v}^{-1}J(r)'\mathrm{d}r$. Because we have approximated $W_v(s) = [W_{v,1}(s)', \ldots, W_{v,N}(s)']'$ by random walks of length 1,000 we take these $1,000m$ (or $1,000s$ in the first integral) discrete points as sampling points for an approximation of the integral by Riemann sums. To this end, we replace all (co)variance terms by consistent estimator based on the calibration period, i.e., we replace Ω_{vv} by $\hat{\Omega}_{vv;m}$ and $\Omega_{u\cdot v}$ by $\hat{\Omega}_{u\cdot v;m}$. More precisely, with $\mathcal{W}_v(s) = [\mathcal{W}_{v,1}(s)', \ldots, \mathcal{W}_{v,N}(s)']$, $\mathcal{X}_{j,v} = [\mathcal{X}_{j,v,1}', \ldots, \mathcal{X}_{j,v,N}']$ and $\hat{\Omega}_{vv;m}^{1/2}\mathcal{W}_v(s) =: \hat{\mathcal{B}}_v(s) = [\hat{\mathcal{B}}_{v,1}(s)', \ldots, \hat{\mathcal{B}}_{v,N}(s)']'$ we use

$$T^{-1}\sum_{r=1}^{[sT]}\hat{\mathcal{B}}_v(r/T) = T^{-1}\sum_{r=1}^{[sT]}\hat{\Omega}_{vv;m}^{1/2}\mathcal{W}_v(r/T) = T^{-1}\sum_{r=1}^{[sT]}T^{-1/2}\sum_{j=1}^{r}\hat{\Omega}_{vv;m}^{1/2}\mathcal{X}_{j,v} \Rightarrow \int_0^s B_v(r)\mathrm{d}r, \tag{1.65}$$

as $T \to \infty$ and

$$\begin{aligned} &T^{-1}\sum_{r=1}^{[mT]}\begin{bmatrix} D(r/T) & \ldots & D(r/T) \\ \hat{\mathcal{B}}_{v,1}(r/T) & \ldots & \hat{\mathcal{B}}_{v,N}(r/T)\end{bmatrix}\hat{\Omega}_{u\cdot v;m}^{-1}\begin{bmatrix} D(r/T) & \ldots & D(r/T) \\ \hat{\mathcal{B}}_{v,1}(r/T) & \ldots & \hat{\mathcal{B}}_{v,N}(r/T)\end{bmatrix}' \\ &\Rightarrow \int_0^m J(r)\Omega_{u\cdot v}^{-1}J(r)'\mathrm{d}r, \end{aligned} \tag{1.66}$$

as $T \to \infty$ and again argue for $T = 1,000$ being large enough for a satisfying approximation. For the third integral define $\mathcal{W}_{u\cdot v}(s) := [\mathcal{W}_{u\cdot v,1}(s), \ldots, \mathcal{W}_{u\cdot v,N}(s)]'$ and $\mathcal{W}_{u\cdot v,n}(s) := T^{-1/2}\sum_{j=1}^{[sT]}\mathcal{X}_{j,u\cdot v,n}$, where $\mathcal{X}_{j,u\cdot v,n}$ are i.i.d. standard normal random variables. Then, $\mathcal{W}_{u\cdot v}(s)$ converges weakly to a vector of independent standard Brownian motions $W_{u\cdot v}(s)$. By the definition of the Itō-Integral we have

$$\begin{aligned} &\sum_{r=1}^{[mT]}\begin{bmatrix} D(r/T) & \ldots & D(r/T) \\ \hat{\mathcal{B}}_{v,1}(r/T) & \ldots & \hat{\mathcal{B}}_{v,N}(r/T)\end{bmatrix}\hat{\Omega}_{u\cdot v;m}^{-1/2}\{\mathcal{W}_{u\cdot v}(r/T) - \mathcal{W}_{u\cdot v}((r-1)/T)\} \\ &\Rightarrow \int_0^m J(r)\Omega_{u\cdot v}^{-1/2}\mathrm{d}W_{u\cdot v}(r) \end{aligned} \tag{1.67}$$

for $T \to \infty$. In case of PFM-OLS estimation (Section 1.2.1) the limit process does not depend on (co)variance terms (see Lemma 3) except the conditional long-run covariance $\omega_{u\cdot v}^2$ that cancels out due to self-normalization of the detectors. Thus, the (co)variance matrices are left out in the simulation of critical values and convergence to $W_v(s)$ instead of $B_v(s)$ in (1.65) is the result. Additionally replacing $\hat{\mathcal{B}}_{v,n}(s)$ by $\mathcal{W}_{v,n}(s)$ yields convergence to $\int_0^m J^W(r)J^W(r)'\mathrm{d}r$ in (1.66) and $\int_0^m J^W(r)\mathrm{d}W_{u\cdot v}(r)$ in (1.67). In order to handle FM-SUR estimation from Section 1.2.3 we need to ensure convergence to $\mathbf{J}(r)$ instead of $J(r)$ in all limits above. To this end, we replace

$$\begin{bmatrix} D(r/T) & \dots & D(r/T) \\ \hat{\mathcal{B}}_{v,1}(r/T) & \dots & \hat{\mathcal{B}}_{v,N}(r/T) \end{bmatrix} \quad \text{by} \quad \operatorname{diag}\left(\begin{bmatrix} D(r/T) \\ \hat{\mathcal{B}}_{v,1}(r/T) \end{bmatrix}, \dots, \begin{bmatrix} D(r/T) \\ \hat{\mathcal{B}}_{v,N}(r/T) \end{bmatrix}\right)$$

in equations (1.66) and (1.67).

Using numerical integration it is easy to approximate integrals of $\mathbf{W_{u\cdot v}}(s)$ and, hence, any of the limiting distributions of the detectors $\mathcal{H}(s)$ by, say, $\mathcal{H}_{\text{approx}}(s)$. Computing $\max_{s=-[-mT],\dots,T}\{\frac{\mathcal{H}_{\text{approx}}(s)}{g(s)}\}$ generates one simulated observation of the monitoring statistic. Replicating this, for example, 1,000,000 times we approximate the distribution of $\sup_{m\leq s\leq 1}\{\frac{\mathcal{H}(s)}{g(s)}\}$ and store the $90.0\%, 90.1\%, \dots, 99.9\%$ quantiles. For both, FM-SUR and PFM-GLS estimation, the limiting distribution depends on the covariance structure, thus, critical values cannot be tabulated in general and have to be simulated for each application of the test seperately. Therefore, the number of replications might be reduced depending on available computational power. If we use PFM-OLS estimation the limiting distribution of $\mathcal{H}(s)$ is independent of the long-run covariance structure for all detectors covered here. Consequently, we are able to tabulate critical values.

1.C. Longrun Covariance and Correlation Matrices

In this section we display the longrun covariance and longrun correlation matrices of the three application examples as well as the respective matrices of the data generating process used in Section 1.3.3 and in a robustness check of the detectors based on PFM-OLS against violations of Assumption 4. Figure 1.9 displays an excerpt of the robustness check for PFM-OLS based detectors and small N. The data generating process is the same as in Section 1.3.3 for the PFM-GLS and FM-SUR based detectors and the results show that the PFM-OLS based detectors have reaonable empirical size here.

For our choice of $\rho_1 = \rho_2 = 0.3$ and $\tilde{\rho} = 0.9$ the longrun correlation matrix of the errors $\eta_t = [u_{1,t}, u_{2,t}, v_{1,t,1}, v_{1,t,2}, v_{2,t,1}, v_{2,t,2}]'$ in the data generating process (1.42) and (1.43)

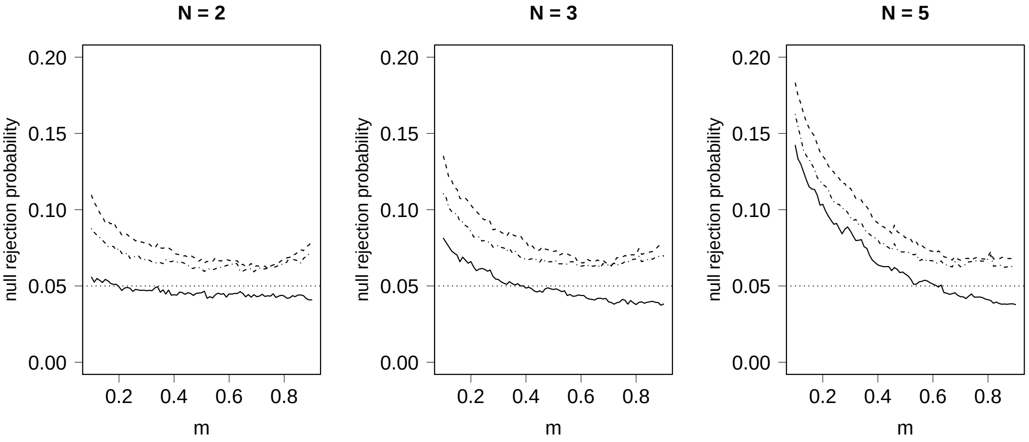

Figure 1.9.: Null rejection probability in correlated homogeneous cointegrating regressions (Section 1.3.3) with $T = 500$, $\rho_1 = \rho_2 = 0.3$, $\tilde{\rho} = 0.9$ and PFM-OLS estimation. The lines represent $\hat{H}^{m,+}_{1,\mathrm{PFM}}$ (solid), $\hat{H}^{m,+}_{2,\mathrm{PFM}}$ (dashed) and $\hat{H}^{m,+}_{3,\mathrm{PFM}}$ (dotdashed).

is

$$\mathrm{Corr}(\eta_t) = \begin{pmatrix} 1.000 & 0.241 & 0.492 & 0.492 & 0.466 & 0.466 \\ 0.241 & 1.000 & 0.466 & 0.466 & 0.492 & 0.492 \\ 0.492 & 0.466 & 1.000 & 0.900 & 0.900 & 0.900 \\ 0.492 & 0.466 & 0.900 & 1.000 & 0.900 & 0.900 \\ 0.466 & 0.492 & 0.900 & 0.900 & 1.000 & 0.900 \\ 0.466 & 0.492 & 0.900 & 0.900 & 0.900 & 1.000 \end{pmatrix}. \tag{1.68}$$

and the estimated longrun correlation matrices in the three application examples are

$$\widehat{\mathrm{Corr}}(\eta_t) = \begin{pmatrix} 1.000 & 0.508 & 0.074 & -0.093 & 0.088 & -0.074 \\ 0.508 & 1.000 & 0.130 & -0.157 & 0.143 & -0.136 \\ 0.074 & 0.130 & 1.000 & -0.948 & 0.980 & -0.947 \\ -0.093 & -0.157 & -0.948 & 1.000 & -0.945 & 0.910 \\ 0.088 & 0.143 & 0.980 & -0.945 & 1.000 & -0.938 \\ -0.074 & -0.136 & -0.947 & 0.910 & -0.938 & 1.000 \end{pmatrix} \text{ for } \begin{pmatrix} \text{USD-XBT-SEK} \\ \text{EUR-XBT-GBP} \end{pmatrix}, \tag{1.69}$$

$$\widehat{\mathrm{Corr}}(\eta_t) = \begin{pmatrix} 1.000 & 0.818 & -0.039 & -0.010 & -0.041 & 0.016 \\ 0.818 & 1.000 & -0.019 & 0.023 & -0.043 & -0.019 \\ -0.039 & -0.019 & 1.000 & -0.926 & 0.900 & -0.965 \\ -0.010 & 0.023 & -0.926 & 1.000 & -0.869 & 0.905 \\ -0.041 & -0.043 & 0.900 & -0.869 & 1.000 & -0.880 \\ 0.016 & -0.019 & -0.965 & 0.905 & -0.880 & 1.000 \end{pmatrix} \text{ for } \begin{pmatrix} \text{USD-XBT-CAD} \\ \text{GBP-XBT-EUR} \end{pmatrix}, \tag{1.70}$$

and

$$\widehat{\mathrm{Corr}}(\eta_t) = \begin{pmatrix} 1.000 & 0.768 & 0.018 & 0.037 & -0.052 & 0.011 \\ 0.768 & 1.000 & 0.058 & -0.024 & -0.064 & -0.025 \\ 0.018 & 0.058 & 1.000 & -0.975 & 0.958 & -0.991 \\ 0.037 & -0.024 & -0.975 & 1.000 & -0.961 & 0.976 \\ -0.052 & -0.064 & 0.958 & -0.961 & 1.000 & -0.959 \\ 0.011 & -0.025 & -0.991 & 0.976 & -0.959 & 1.000 \end{pmatrix} \text{ for } \begin{pmatrix} \text{USD-XBT-AUD} \\ \text{RUB-XBT-EUR} \end{pmatrix}. \tag{1.71}$$

With a χ^2-test in the style of Breusch and Pagan (1980), it is possible to test for cross-sectional independence of the systems. The null hypothesis is $H_0 : \mathrm{Corr}(u_{1,t}, u_{2,t}) = 0$ and the test statistic is $T_0 \cdot \widehat{\mathrm{Corr}}(u_{1,t}, u_{2,t})$, where $T_0 = 134$ is the length of the calibration sample. The test statistic is asymptotically χ^2_1-distributed under the null hypothesis and in all three cases, the p-value of the test is smaller than 10^{-8}.

The long-run covariance matrix of the errors η_t in the data generating process (1.42) and (1.43) in Section 1.3.3 for $\rho_1 = \rho_2 = 0.3$ and $\tilde{\rho} = 0.9$ is

$$\mathrm{Cov}(\eta_t) = \begin{pmatrix} 22.759 & 1.984 & 4.506 & 4.506 & 4.269 & 4.269 \\ 1.984 & 22.759 & 4.269 & 4.269 & 4.506 & 4.506 \\ 4.506 & 4.269 & 3.063 & 2.756 & 2.756 & 2.756 \\ 4.506 & 4.269 & 2.756 & 3.063 & 2.756 & 2.756 \\ 4.269 & 4.506 & 2.756 & 2.756 & 3.063 & 2.756 \\ 4.269 & 4.506 & 2.756 & 2.756 & 2.756 & 3.063 \end{pmatrix} \times 10^{-3}. \tag{1.72}$$

and the estimated long-run covariance matrices in the three application examples are

$$\widehat{\mathrm{Cov}}(\eta_t) = \begin{pmatrix} 39.803 & 49.693 & 1.076 & -1.389 & 1.293 & -1.108 \\ 49.693 & 240.68 & 4.641 & -5.761 & 5.164 & -4.975 \\ 1.076 & 4.641 & 5.260 & -5.128 & 5.215 & -5.126 \\ -1.389 & -5.761 & -5.128 & 5.560 & -5.172 & 5.063 \\ 1.293 & 5.164 & 5.215 & -5.172 & 5.382 & -5.133 \\ -1.108 & -4.975 & -5.126 & 5.063 & -5.133 & 5.565 \end{pmatrix} \times 10^{-3} \text{ for } \begin{pmatrix} \text{USD-XBT-SEK} \\ \text{EUR-XBT-GBP} \end{pmatrix}, \tag{1.73}$$

$$\widehat{\mathrm{Cov}}(\eta_t) = \begin{pmatrix} 40.111 & 25.945 & -0.536 & -0.134 & -0.615 & 0.220 \\ 25.945 & 25.105 & -0.207 & 0.245 & -0.508 & -0.208 \\ -0.536 & -0.207 & 4.683 & -4.287 & 4.557 & -4.650 \\ -0.134 & 0.245 & -4.287 & 4.574 & -4.349 & 4.313 \\ -0.615 & -0.508 & 4.557 & -4.349 & 5.471 & -4.586 \\ 0.220 & -0.208 & -4.650 & 4.313 & -4.586 & 4.961 \end{pmatrix} \times 10^{-3} \text{ for } \begin{pmatrix} \text{USD-XBT-CAD} \\ \text{GBP-XBT-EUR} \end{pmatrix}, \tag{1.74}$$

and

$$\widehat{\mathrm{Cov}}(\eta_t) = \begin{pmatrix} 28.007 & 83.582 & 0.231 & 0.492 & -0.690 & 0.145 \\ 83.582 & 423.098 & 2.995 & -1.222 & -3.301 & -1.255 \\ 0.231 & 2.995 & 6.201 & -6.066 & 6.026 & -6.083 \\ 0.492 & -1.222 & -6.066 & 6.237 & -6.063 & 6.006 \\ -0.690 & -3.301 & 6.026 & -6.063 & 6.378 & -5.972 \\ 0.145 & -1.255 & -6.083 & 6.006 & -5.972 & 6.078 \end{pmatrix} \times 10^{-3} \text{ for } \begin{pmatrix} \text{USD-XBT-AUD} \\ \text{RUB-XBT-EUR} \end{pmatrix}. \tag{1.75}$$

The difference in estimated long-run variances of $u_{1,t}$ and $u_{2,t}$ in each of the currency triplets (USD-XBT-SEK)-(EUR-XBT-GBP) and (USD-XBT-AUD)-(RUB-XBT-EUR) does not correspond to the simplification proposed in Section 1.3.3 but the number of estimated parameters remains reasonably low without any simplification as $N = k = 2$ are small.

Chapter 2.

Reference Class Selection in Similarity-Based Forecasting of Corporate Sales Growth

2.1. Introduction

The forecasting of future cashflows and an appropriate discount rate is pivotal for the valuation of companies and active management of equity investments (e.g., Guerard et al., 2015, in portfolio construction). In order to tackle this task, analysts have to forecast performance indicators like corporate sales or operating margins for different periods of time. However, in general there is low predictability of growth rates (see Chan et al., 2003) and forecasts are often based on heuristics and were empirically shown to be biased as well as overoptimistic (see, e.g., Tversky and Kahneman, 1973, 1974; Kahneman and Tversky, 1973; Cooper et al., 1988). In our context, survey results of Kunte (2015) among financial market practitioners show that herding (34%), confirmation (20%), overconfidence (17%), availability (15%) and loss aversion (13%) are the behavioral biases that affect investment decisions the most. Lim (2001) reviews analysts' bias, Jones and Johnstone (2012) find proof for overoptimism while Löffler (1998) unravels overconfidence and underreaction to news and Lee et al. (2008) identify negligence of buisness cycles as a source of bias. Ashton and Cianci (2007) discuss differences between buy-side and sell-side analysts' forecasts and Stotz and von Nitzsch (2005) analyze reasons for analysts' overconfidence.

A large part of the distorted forecasts is due to the fact that forecasts are often solely based on the so called *inside view*, which considers each forecasting challenge as unique and neglects statistical information, as well as results of similar forecast challenges (Kahneman and Lovallo, 1993). Thus, it can be very helpful to use empirical data and existing experience, the so called *outside view*, in order to identify and reduce the aforementioned biases (Tetlock and Gardner, 2016). The basic idea of the outside view is the definition of a reference class which includes objects of comparison similar to the initial object (Kahneman and Tversky, 1979; Lovallo and Kahneman, 2003). By means of this objective data set the forecaster becomes empowered to challenge and improve his forecast (Kahneman and Tversky, 1979). Adjusting or correcting forecasts is an already established tool in the financial and forecasting literature in terms of judgementally adjusting model based forecasts by experts (Wolfe and Flores, 1990; Sanders and Ritzman, 2001; De Bruijn and Franses, 2017), combining statistical forecasts with analysts' predictions (Lobo, 1991; Bunn and Wright, 1991) and combining analysts' forecasts or using consensus forecasts (Butler and Saraoglu, 1999; Ramnath et al., 2005; Jame et al., 2016). However, Du and McEnroe (2011) examine reports by research firms with multiple analysts' forecasts. Similar forecasts leads to overconfidence while highly varying forecasts diminish confidence. Further, Du and Budescu (2018) show that the hit rates of analysts for earnings per share in 2014 range from 37% to 52%, depending on the forecast horizon. Our contribution will add to the toolbox of analysts and investors by the property to directly calculate prediction intervals.

The concepts of the outside view and reference classes are well known in literature and practice, e.g., in infrastructure projects (Flyvbjerg, 2006, 2008; Themsen, 2019) or software development (Shmueli et al., 2016). Moreover, the use of base rates, i.e., distributional information, is recommended by Armstrong (2005) and is part of professional forecasters and analysts' training (Tetlock and Gardner, 2016) which is shown to improve their performance (Chang et al., 2016). Especially Karvetski et al. (2021) show that the use of base rates has a positive effect on forecast accuracy but in general there has been paid more attention to the biases than to debiasing (Chang et al., 2016). Green and Armstrong (2007) describe a procedure to include analogies in the forecasting process and Lovallo et al. (2012) conduct an empirical study using the outside view to forecast stock returns but both suffer from a subjective choice of similar objects such that the resulting reference classes are prone to the availability bias described by Tversky and Kahneman (1973). Noteworthy, Knudsen et al. (2017) construct peer groups of comparable companies for corporate valuation objectively by using a measure of similarity but these reference classes consist of only six elements elevating the probability of bias again. Surprisingly, there is a lack of studies which investigate how to construct optimal reference classes for the forecasting of future cash flows and the related performance indicators. To the best of our knowledge, the only existing concept is proposed by Mauboussin and Callahan (2015). They define 11 reference classes based on the size of the actual sales level in order to derive base rates for the growth rate of sales. However, the defined reference classes are neither theoretically derived nor empirically backtested. Thus, the quality of the reference classes and the added value for the analysts remain vague.

This chapter fills the previously mentioned gap in literature. On the one hand, we propose a method to find appropriate outside views for sales forecasts of analysts in Section 2.2. Hence, we define reference classes for each analyzed company separately by means of additional companies that share similarities to the firm of interest with respect to a specific co-variate, here called reference variable. This approach is easy to implement and interpret as we deliberately restrict the analysis to exactly one reference variable at once, which also ensures that only a parsimonious amount of data is required. Thus, the proposed method is well suited for practical applications. On the other hand, we describe a data set consisting of 21,808 US firms over the time period from 1950 to 2019 in Section 2.3 that we use to evaluate different reference variables. To this end, we backtest their quality in Section 2.4 by means of goodness-of-fit tests and by calculating a novel measure Δ_q based on predicted quantiles of probability integral transform values. This analysis yields that in particular the past operating margins are good reference variables for the distribution of future sales. Moreover, a case study in Section 2.5 compares our forecasts to actual analysts' estimates in order to show the practical usefulness and demonstrates how to apply the results of our approach. Lastly, Section 2.6 concludes.

2.2. Reference Class Selection

The notion of reference class forecasting is based on ideas of Princeton psychologist and Nobel prize winner Daniel Kahneman and his co-author Amos Tversky. It originates in theories of planning and decision-making under uncertainties and is motivated by the fact that forecasts are often based on heuristics and were empirically shown to be biased as well as overoptimistic. In order to overcome this issue it is advisable to contrast the inside view, that is, information on the specific case at hand, with the outside view, that is, information on a class of similar cases. This may include, for example, statistical or empirical distributional information as well as base rates and is a promising approach to overcome overoptimism, wishful thinking or strategic misrepresentations.

Kahneman and Tversky (1979) have introduced a corrective procedure for biases of predictions which involves five steps. First, the forecaster has to identify a set of similar cases which define the reference class and provide the distribution of outcomes to be predicted. This distribution has either to be assessed directly or to be estimated within the next step. At this point the expert uses their available information on the case for an inside prediction. In the fourth step the expert needs to assess the predictability of their forecasts. In case of linear prediction, this may be the correlation between their predictions and the outcomes. Finally, the inside prediction is corrected and adjusted towards the mean of the reference class.

While each of the five steps has its own pitfalls in practice, we focus on the first one and provide guidance on how to select an appropriate reference class. This is of major importance as Kahneman and Tversky (1979) gave no guideline how to build reference classes apart from the general rule to use similar cases. Moreover, there is a fundamental conflict of objectives in defining the reference class. On the one hand, it would be desirable to take as many cases into account as possible. However, it is crucial that heterogeneity does not become too large and each object is still comparable to the initial one. On the other hand, each element within the reference class should be similar to the initial object, whereby the risk arises that the class becomes too small and the objects too similar. In this case the probability of a biased forecast is elevated again. Based on this fact Lovallo and Kahneman (2003) state: *"Identifying the right reference class involves both art and science."*

In literature, there are several studies dealing with reference class building. For example, Lovallo et al. (2012) report two case studies with respect to private-equity investment decisions and film revenue forecasts. However, and to the best of our knowledge, there is a gap with respect to reference classes for the forecasting of future cash flows and the related performance indicators. The only existing concept is proposed by Mauboussin and Callahan (2015). They state that sales growth is the most important driver of corporate value and define reference classes by sorting the firms' real sales in 10 deciles as well as an 11th class for the top one percentile. To this end they use historical data of the S&P1500 from 1994-2014. In total they show the distribution of growth rates for 55 reference classes (11 size ranges multiplied by five time horizons) but give neither a theoretical justification for nor an empirical backtest of their proposed procedure. Thus, the quality of the proposed reference classes and the added value for the analysts remain open questions, especially as they use clustered data which has a substantial problem in general. As an example, Figure 2.1 shows three clusters constructed by the k-means algorithm for a simulated data cloud and highlights the pitfall that an element on the border of one cluster may be closer to the elements of another cluster than to the majority of elements in its own cluster – a general drawback of procedures using cluster algorithms.

In order to overcome this drawback we will present an alternative method which does

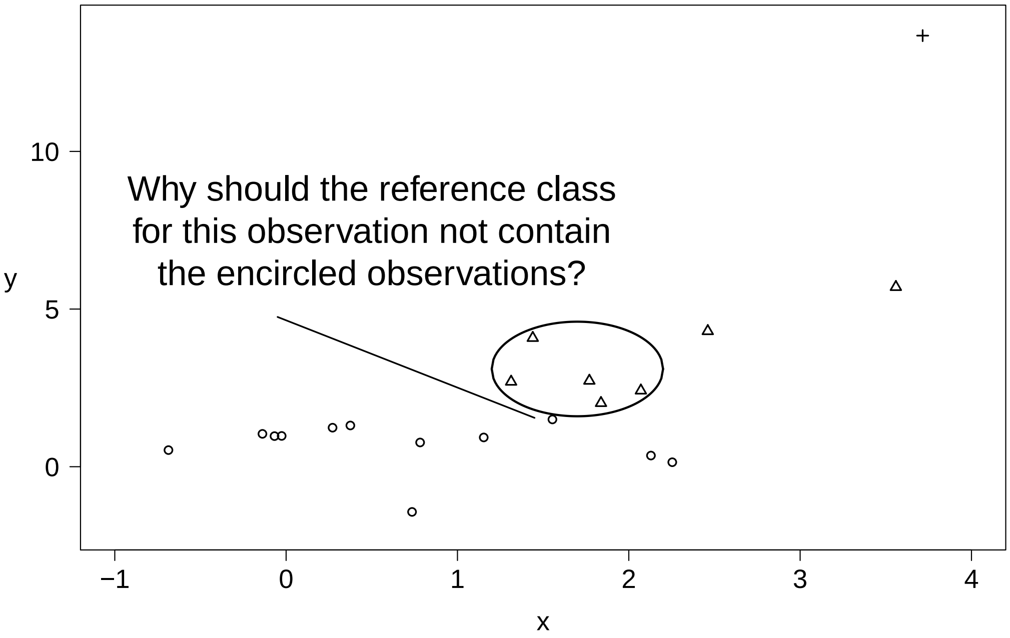

Figure 2.1.: These three clusters constructed by the k-means algorithm for a simulated data cloud highlight the risk that elements on the border of one cluster may be closer to the elements of another cluster than to the majority of elements in their own clusters.

not rely on cluster algorithms and finds reference classes for each analyzed company separately whereby the approach is easy to implement and interpret. Moreover, we evaluate the resulting reference classes out-of-sample on a data set ranging from 1950 to 2019 in order to be able to make a meaningful quality valuation. The following two subsections provide the theoretical foundations.

2.2.1. Theoretical Framework

We aim to forecast $Y_{i,t+h}$, that is, an h-step ahead forecast of the random variable $\{Y_{i,t}\}$ for firm i at time t. In the following applications this is sales growth but basically it could be any other quantity of interest. At this point we assume that a sufficient amount of historical data of additional firms is available in order to assess the distribution of

$Y_{i,t+h}$. We base the reference class on a specific reference variable $\{X_{i,t}\}$.[1] The idea is now to build a reference class R by finding firms j in the past which are similar to firm i with respect to the reference characteristic and in some norm $||\cdot||$, that means,

$$||\{X_{i,t}\} - \{X_{j,s}\}||$$

shall be *small*, where $s + h \leq t$ ensures that the realization of $Y_{j,s+h}$ is available. For example, we could use all companies which had an operating margin ±1 percentage points in comparison to the actual margin of firm i during the last 10 years. Figure 2.1 illustrates the difference of our approach to a classical cluster analysis. We do not try to find disjoint clusters of firms, but aim at finding neighbors for each firm separately. A forecast for the distribution of $Y_{i,t+h}$, which is used as an outside view, is now given by the empirical distribution of the values $Y_{j,s+h}$, $(j, s) \in R$.

The first assumption behind the approach is the existence of a *market mechanism*, say a smooth function f_h such that $Y_{i,t+h} \sim f_h(\{X_{i,t}\})$. Moreover, we need some kind of stationarity assumption so that this mechanism works similarly over time and we have $Y_{j,s+h} \sim f_h(\{X_{j,s}\})$, $(j, s) \in R$, for the outcomes within the reference class. If $\{X_{i,t}\}$ is close to $\{X_{j,s}\}$, which is supposed to be provided by finding suitable reference classes, $f_h(\{X_{i,t}\})$ is close to $f_h(\{X_{j,s}\})$ and the empirical distribution function of $Y_{j,s+h}$ is a good approximation for the distribution of $Y_{i,t+h}$. Note, the goal of this study is not to get information about f_h, but to get information about how suitable reference classes are.

2.2.2. Performance of Procedure

By means of the resulting distributional information we can assess predictions (e.g., by experts or analysts or model based forecasts) or we can assess the suitability of the reference class as a distributional forecast. To this end, a direct comparison to the outcome is not possble as, e.g., in the case of point forecasts. Here, we evaluate the empirical cumulative distribution function of the reference class at the (known) realization,

[1] The reference variables are also called reference characteristics or predictor variables in Kahneman and Tversky (1979) and Theising et al. (2023) but we stick to the term ‘reference variable’ as they are random variables here, used for reference class selection and do not predict directly but only implicitly through the selection.

that is, we calculate

$$\mathbb{P}(Y_{i,t+h} \leq y_{i,t+h}) \approx n^{-1} \sum_{(j,s)\in R} \mathbb{1}\{Y_{j,s+h} \leq y_{i,t+h}\}, \tag{2.1}$$

where $n = |R|$. Repeating this for multiple firms and points in time results in a sample of size m, whereas the values lie in the interval $[0, 1]$. If the approximation of the distribution is valid, (2.1) is roughly the probability integral transform and consequently we approximately have realizations from a uniform distribution on $[0, 1]$. To assess the forecast ability of the different reference variables, we consider measures that determine how close this approximation is. This is done with classical statistical goodness-of-fit tests as well as a comparison of quantiles.

Let G_m be the empirical distribution function of these frequencies $\{p_k\}_{k=1,\ldots,m}$ and let G be the true distribution function of the counterparts of these frequencies in the population. Let G_0 be the distribution function of the uniform distribution on $[0, 1]$. The considered pair of hypotheses is $H_0 : G = G_0$ vs. $H_1 : G \neq G_0$ and the corresponding two test statistics are given by $\sqrt{m}\sup_{x\in[0,1]} |G_m(x) - G_0(x)|$ (Kolmogorov-Smirnov) and $m \int_0^1 [G_m(x) - G_0(x)]^2 \mathrm{d}G_0(x)$ (Cramer-von-Mises).

However, we do not consider the actual tests' decisions. Working with sample sizes between $100,000$ and $300,000$, depending on hyper parameters, we face the problem pointed out by Berkson (1938): "Any consistent test will detect any arbitrary small change in the [distribution] if the sample size is sufficiently large". Thus, most p-values would be very small or even get reported as 0 by software. Avoiding this problem, we focus on the value of the test statistics and rank the different combinations of reference variable and hyper parameters based on these values.

A third and new measure of ranking the models consists of comparing the quantiles of probability integral transform values. This means that for a finite number of quantile levels, we consider the absolute difference Δ_q between the quantiles of $\{p_k\}_{k=1,\ldots,m}$ and the quantiles of the uniform distribution on $[0, 1]$. These differences are summed up and ranked.

2.3. Data Set

In order to find the best reference variable and appropriate hyper parameters we analyze their performance on an historic data set with regards to finding optimal refer-

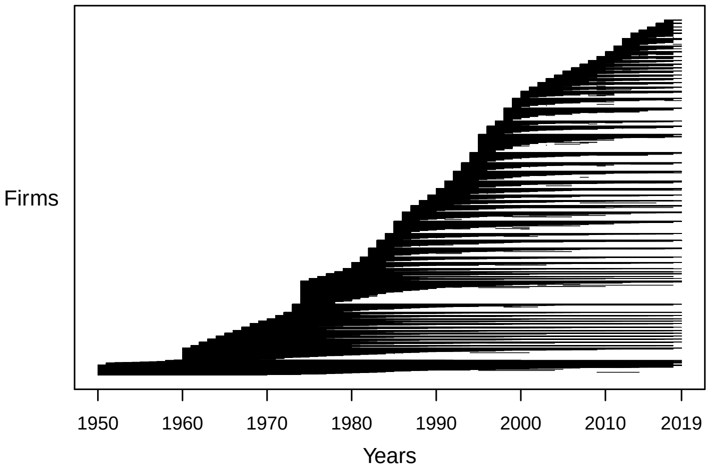

Figure 2.2.: This figure displays the time series properties of firms. Each horizontal line represents one of the 21,808 firms ordered from bottom to top by three criteria: 1. the first year of appearance in the data set, 2. the number of observations of the firm, 3. the number of consecutive observations of the firm.

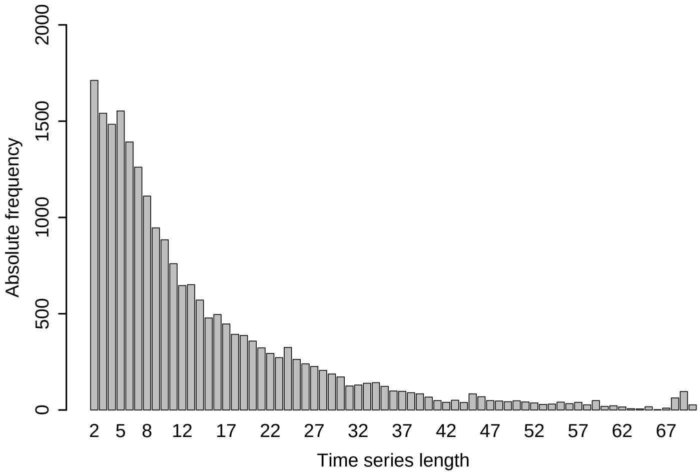

Figure 2.3.: This barplot shows the number of observations per firm in the data set, that is the empirical distribution of time series length.

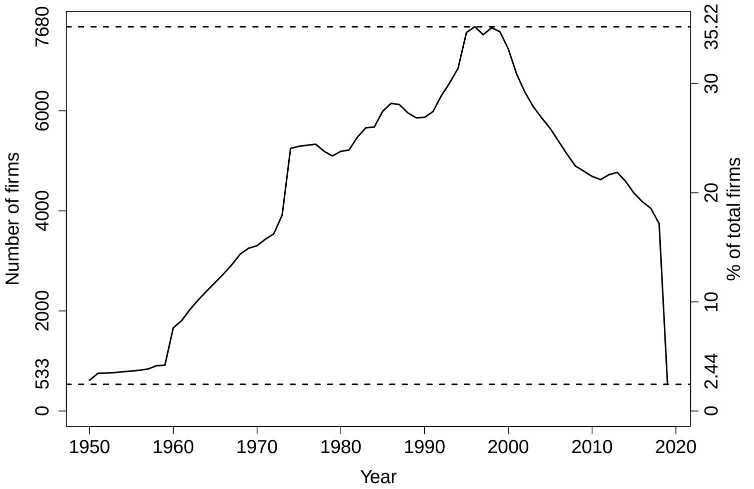

Figure 2.4.: Number of companies over time. The left vertical axis shows the number of firms (i.e., observations) per year and the right vertical axis covers the number of firms as a proportion of the total number of firms.

ence classes. We use Compustat North America fundamentals annual data from 1950 to 2019 by S&P Global Market Intelligence (2020) and limit our analysis to US firms excluding companies from the financial and real-estate sector. Firms without sales information or only one observation are discarded due to our interest in predicting distributions of sales growth. We merge these data with stock-exchange information from the Center for Research in Security Prices (CRSP, 2020) daily stock of the University of Chicago Booth School of Business. All variables collected in US dollar are inflation adjusted to 1982 – 1984 US dollar using monthly inflation rate data from the consumer price index for all urban consumers (all items in US city average) by the U.S. Bureau of Labor Statistics (2020).

The data set consists of 303,628 observations on 21,808 firms with CRSP stock exchange market information on 206,221 observations of 17,099 firms in total. The length of the time series of the different firms varies considerably (c.f. Figures 2.2 and 2.3) as well as the number of observations per year (c.f. Figure 2.4). To put this in perspective, there is an influence of survivorship in the data set. Our later backtest focusses on one-, three-, five- and 10-year predictions and the survivorship rates are 97.25% for one year, 89.61% for three years, 76.12% for five years and 48.20% for 10 years.

We select and investigate the most common metrics used for fundamental analysis as possible reference variables whereby some of them relate to the company directly while

Table 2.1.: Description of reference variables. EBIT is earning before interest and taxes, market cap. is market capitalization and pp is percentage points. A variable summary can be found in Table 2.2.

Abbreviation	Reference Variables	Description
at	total assets	in million USD
opmar	operating margin	EBIT divided by sales (in %)
–	sales	in million USD
seq	shareholder equity	total assets minus total liabilities (in million USD)
–	major group	first two digits of SIC, 63 groups
–	industry group	first three digits of SIC, 250 groups
–	β	slope of regressing daily return on market return
P/B	price-to-book ratio	market cap. divided by shareholder equity
P/E	price-to-earnings ratio	market cap. divided by net income
salesGR$_\tau$	τ-year past sales growth	current sales divided by sales τ years ago (in %)
opmarΔ_τ	τ-year past operating margin delta	current operating margin minus operating margin τ years ago (in pp)

some others are market parameters. To be more precise, observed key figures for all companies are sales, operating margin, total assets, shareholder equity, the SIC (standard industrial classification), β, the price-to-earnings ratio and the price-to-book ratio. Using sales and operating margin information over time, we construct one- to 10-year past sales growth and one- to 10-year past operating margin delta as additional possible reference variables where the necessary data are available. Instead of SIC itself, we derive a firm's major and industry group and use these groups to construct reference classes as a benchmark of the typical current practice. In Table 2.1 we provide a description and in Table 2.2 a summary of the reference variables used to construct reference classes including relevant quantiles, their means and the number of missing values in the data set.

We aim to forecast distributions of future sales growth while using exactly one of the reference variables to construct reference classes. To be more precise, we construct one-, three-, five- and 10-year future sales growth forecasts using temporal information in the data set. Table 2.3 displays the base rates for the full universe of data, i.e., the historical sales compound annual growth rate (CAGR). Here, the tails of the distribution get lighter, the (2.5%-trimmed) standard deviation declines, the (2.5%-trimmed) mean gets closer to the median and the distribution more centered the longer the forecast horizon is, as it is visible in Figure 2.5 as well. By a 2.5%-trimmed mean or standard deviation we are referring to the arithmetic mean or standard deviation, respectively, where the largest

Table 2.2.: Summary of reference variables (ref.var.) as in Table 2.1, where qu. is quantile. salesGR is compound annual growth rate in % in this table and opmarΔ is the mean annual past operating margin delta in this table to simplify comparison across lags. The summary on major and industry groups covers the group sizes.

Ref. Var.	2.5% qu.	25% qu.	Median	Mean	75% qu.	97.5% qu.	Missings
at	0.27	11.82	62.31	877.65	337.77	6767.24	2714
opmar	-827.80	-1.19	6.01	-402.68	12.27	34.49	18532
sales	0.00	10.67	67.60	721.10	337.74	5345.51	0
seq	-9.65	3.58	24.00	319.76	128.97	2478.79	19811
major group	10	895	2646	4819.49	5295	25617	0
industry group	38	283	622	1214.51	1248	6793	0
β	-0.28	0.37	0.77	0.83	1.21	2.31	97469
P/B	-6.00	0.59	1.34	2.65	2.57	11.70	100318
P/E	-70.39	-3.45	8.34	11.24	17.69	104.99	98786
salesGR$_1$	-100	-5.39	4.93	115.70	19.24	1465000	31591
salesGR$_2$	-100	-4.18	4.55	17.07	16.33	19090	52164
salesGR$_3$	-100	-3.31	4.32	10.41	14.51	3862	71103
salesGR$_4$	-100	-2.71	4.21	7.90	13.17	1794	88572
salesGR$_5$	-100	-2.22	4.13	6.52	12.23	1019	104702
salesGR$_6$	-100	-1.87	4.05	5.62	11.44	609.50	119372
salesGR$_7$	-100	-1.55	4.00	5.02	10.82	435.80	132772
salesGR$_8$	-100	-1.29	3.98	4.59	10.38	333.90	145044
salesGR$_9$	-100	-1.06	3.95	4.28	9.97	277.10	156300
salesGR$_{10}$	-100	-0.87	3.91	4.03	9.58	205.30	166682
opmarΔ_1	-2824000	-2.73	0.04	-10.15	2.57	2823000	41527
opmarΔ_2	-1412000	-1.96	-0.03	-11.85	1.71	681300	62660
opmarΔ_3	-374800	-1.54	-0.07	4.04	1.26	951200	81829
opmarΔ_4	-326200	-1.27	-0.08	3.89	1.00	691100	99288
opmarΔ_5	-260800	-1.09	-0.08	3.19	0.82	523200	115291
opmarΔ_6	-217300	-0.95	-0.09	0.42	0.69	204400	129585
opmarΔ_7	-107800	-0.84	-0.09	3.81	0.60	185700	142583
opmarΔ_8	-89290	-0.76	-0.08	2.25	0.53	190800	154449
opmarΔ_9	-81610	-0.69	-0.08	3.21	0.46	335300	165288
opmarΔ_{10}	-75350	-0.64	-0.08	3.44	0.41	301700	175265

Table 2.3.: Compound annual sales growth rates for the whole data set. Mean and standard deviation are 2.5% trimmed on both tails, the respective quantiles are in the table.

Full Universe	Base Rates			
CAGR (%)	1-Yr	3-Yr	5-Yr	10-Yr
$\leq$ -25	8.70	5.44	4.00	2.38
]-25,-20]	2.19	1.69	1.28	0.68
]-20,-15]	3.18	2.65	2.13	1.37
]-15,-10]	4.53	4.27	3.71	2.68
]-10,-5]	7.06	7.28	7.11	6.12
]-5,0]	10.92	13.20	14.29	15.64
]0,5]	13.59	17.82	21.17	27.25
]5,10]	11.65	14.33	16.34	20.09
]10,15]	8.24	9.06	9.70	9.95
]15,20]	5.65	5.86	5.77	5.38
]20,25]	4.08	3.95	3.61	2.92
]25,30]	3.05	2.71	2.54	1.76
]30,35]	2.31	2.04	1.73	1.14
]35,40]	1.78	1.54	1.26	0.69
]40,45]	1.46	1.17	0.93	0.48
$>$ 45	11.58	6.99	4.42	1.46
mean	10.62	7.01	5.75	4.62
median	4.93	4.32	4.13	3.91
std	32.30	19.08	14.21	9.20
$q_{0.025}$	-60.01	-44.75	-36.52	-23.91
$q_{0.975}$	206.31	95.19	62.75	35.85

2.5% and the smallest 2.5% of the data are excluded.[2] The (2.5%-trimmed) means of sales CAGR are larger than the respective medians because the growth rates are left bounded and right unbounded and we observe a substantial amount of high values one could characterize as outliers which make the ordinary mean and standard deviation uninformative. In order to restrain the influence of these outliers and to keep the mean and standard deviation informative we use the trimmed versions of these measures. The summary statistic of the sales CAGR can be found in Table 2.1 as the distribution of future and past growth rates in the full data set are identical.

[2] For a vector of sorted observations $\{x_i\}_{i=1,\ldots,n}$ we compute any α-trimmed measure, $0 < \alpha < 1$, based on the trimmed vector of observations $\{x_i\}_{i=[\alpha n]+1,\ldots,n-[\alpha n]}$, where $[\cdot]$ is the floor function.

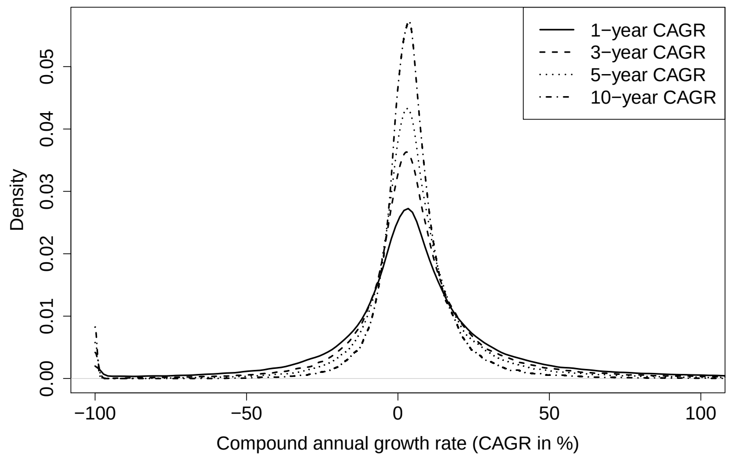

Figure 2.5.: Estimated densities of compound annual sales growth for horizons one, three, five and 10 years. For density estimation on support $[-100, \infty)$ we used the Gaussian kernel with Silverman's rule of thumb as bandwidth.

Table 2.4.: (Hyper) Parameters in single reference variable approach.

Name	Description
reference variable	see Table 2.1
class size	relative size $\in \{0.050, 0.025, 0.010\}$
window	number of past years $\in \{5, 10, 20, 30\}$

2.4. Backtest

By means of a backtest we compare the performance of our new procedure to forecast distributions of sales growth rates to the performance of the benchmark approach by Mauboussin and Callahan (2015) and the typical practice of using industry classifications, here the first two and three digits of SIC, respectively. We include three (hyper) parameters in the backtest where all methods depend on the number of past years to use for reference class construction and only our new procedure depends additionally on the reference variable as well as the size of the reference class (see Table 2.4). Forecast horizons investigated are one, three, five and 10 years.

The parameter window w defines the number of past years to provide candidates of historical observations to construct a reference class. All observations from this window period with known outcomes (i.e., available h-year future sales growth) are candidates for the reference class. In order to backtest out-of-sample, given an initial case firm i at time t, the parameters w and h determine the years of historical data to serve as candidates, namely starting in $t-h-w+1$ and ending in $t-h$ (assuming that at time t all information of the financial year t is available). That means we consider all firms j at times s as candidates for the initial case's reference class, where $t-h-w+1 \leq s \leq t-h$ and the reference variable and h-year sales growth are available.

The size of the reference class, that is, the number of observations it contains, is relative to the number of candidates and defined by the size parameter $c \in (0,1)$ determining which of the candidates $X_{j,s}$ lie closely enough to the initial case $X_{i,t}$ to be a member of the reference class. To be more precise, this means c assesses for which candidate firms j at time s the value $||X_{i,t} - X_{j,s}||$ is considered as *small*. Here, we order the candidates by the reference variable and take the $c/2$ fraction smaller than the initial case's observation and the $c/2$ fraction larger than the initial case's observation. More theoretically, let $\hat{F}_{\text{cand}}$ be the empirical distribution function of all candidates and $\hat{F}_{\text{cand}}^{-1}$ be the associated empirical quantile function of all candidates. Then, all candidate firms j at time s with $|\hat{F}_{\text{cand}}^{-1}(X_{i,t}) - \hat{F}_{\text{cand}}^{-1}(X_{j,s})| \leq c/2$ are chosen as members of the reference class. The parameter c is only relevant for our new approach. To keep the class size constant even if the initial case's reference variable is at the tail of the candidates' distribution, we choose the top or bottom fraction c of the candidates regarding the reference variable if $\hat{F}_{\text{cand}}^{-1}(X_{i,t}) > 1 - c/2$ or $\hat{F}_{\text{cand}}^{-1}(X_{i,t}) < c/2$, respectively. Moreover, the reference class of each case has to consist of at least 20 elements or members in order to make reasonable distribution forecasts and to be considered within our backtest. This requirement applies to the benchmark methods as well.

The benchmark models are the approach of Mauboussin and Callahan (2015) and a simple approach using the major and industry group of a firm and set the bar for our new method. Mauboussin and Callahan (2015) define the reference classes by sorting the candidates' real sales in 10 deciles as well as an 11th class for the top one percentile. We use the major and the industry group in a typical straightforward way to construct a reference class from the set of candidates. In both cases, all candidate firms that are in the same major or industry group, respectively, as the initial case are members of the reference class. Thus, there is no size parameter in either of the benchmark approaches.

Our new approach is analyzed with regards to 27 reference variables, three different class sizes and four different window lengths, thus resulting in 324 different combinations for each forecast horizon. The approach of Mauboussin and Callahan (2015) uses one reference variable and four different window sizes and the typical industry classification approach uses two reference variables and four different window sizes, that are 12 benchmark combinations overall. In total we have 336 different combinations for each forecast horizon.

For each approach and combination of (hyper) parameters we consider each observation in the data set as an initial case, i.e., each firm i at each point in time t (where the firm is in the data set). We construct a reference class if several criteria are met. The reference variable and the full window length of historical data must be available, i.e., $t \geq 1950 + w + h - 1$ since our data set starts in 1950. The h-year future sales growth must be available, so at least $t \leq 2019 - h$. Moreover, firm i must be in the data set at time $t + h$ and the reference class has to consist of at least 20 elements.

After obtaining the reference class for an initial case (i, t) we evaluate the empirical distribution function of the sales growth rates of the reference class elements (base rates) at the realized sales growth rate of firm i at time t. Doing this for all initial cases of a parameter combination provides a sample of forecasted probabilities $\{p_k\}_{k=1,\dots,m}$ of being less or equal to the realized sales growth of the initial case. The sample size m depends on the availability of the reference and forecast variable, the window length and the forecast horizon. If the approximation of the distribution by the reference class is valid we roughly have realizations from a uniform distribution on $[0, 1]$. We then use the Kolmogorov-Smirnov (KS) test statistic and the Cramer-von-Mises (CvM) test statistic to measure the accuracy of the distributional approximation. As a third and novel measure of accuracy, we calculate the differences of the 1%, 5%, 10%, 25%, 50%, 75%, 90%, 95% and 99% quantiles of $\{p_k\}_{k=1,\dots,m}$ and of the uniform distribution on $[0, 1]$, respectively, and sum up the absolute values of these differences to obtain Δ_{q}.

2.4.1. Results of Backtest

Tables 2.5 - 2.8 show an excerpt of our results[3]. We display the best three parameter combinations according to the quantile deviation Δ_{q} and as a comparison the benchmark approach of Mauboussin and Callahan (2015) for the best window length. Moreover, we present the benchmark approaches using industry classification through SIC's major and industry group with the best window length, respectively. The best combinations are in all cases various combinations of the reference variable *past operating margin*

[3] Full results are available upon request.

delta followed next by the reference variable *operating margin* which is why we included the best parameter combination for the operating margin as well. As a comparison to the simpler approach by Mauboussin and Callahan (2015) we also included the best parameter combination for the reference variable *sales*. All reference variables which include only contemporaneous information have the common advantage not to rely on (a lot) of historical information of the initial case.[4] The best parameter combinations all involve a window length of 30 which may be hard to achieve in practice. Hence, we added the best parameter combinations for window lengths five and 10 to get an impression of the influence of historical information. Thus, we report 10 results for each forecast horizon except for one-year sales growth. Here, the best parameter combination for window length 10 and the best parameter combination for reference variable *operating margin* coincide.

In order to get a sense of the measure Δ_q, we consider the best reference variable *six-year operating margin delta* for forecasting one-year ahead sales growth from Table 2.5. Here we have $\Delta_\mathrm{q} = 0.0155$, which is the sum of the absolute quantile deviations for nine quantiles. So, the mean absolute deviation of these quantiles is 0.17 percentage points. Therefore, the backtest shows that we miss the quantile levels of the underlying distribution of one-year ahead sales growth on historical data by only 0.17 percentage points on average. Assuming that a practitioner constructs, for example, a 95% prediction interval from the reference class the error in coverage rate should be negligible.

The results are consistent across the accuracy measures and the relative class size does not influence the results substantially. All goodness-of-fit measures generally improve with a shorter forecast horizon. The *past operating margin deltas* are the best reference variables using a window of length 30. In contrast, the best reference variables for window lengths of five and 10 are the *operating margin* for forecast horizons one and three while the *price-to-earnings ratio* is best for the forecast horizon five. For forecast horizon 10 *price-to-earnings ratio* is optimal for the window length five and the *10-year past operating margin delta* for a window length of 10.

Constructing reference classes by the benchmark procedure using *major or industry groups* yields the worst results for horizons one, three and five. Only for a 10-year horizon the industry classification by groups results in more accurate distributional forecasts.

[4] The necessity of historical information to use the past operating margin deltas as reference variables reduces the amount of data and produces the risk of survivorship bias causing the better accuracy. We performed a robustness check where we limited the data set for each forecast horizon to the observations with available best reference variable of this backtest. The past operating margin deltas still performed best. Results are available upon request.

Table 2.5.: Comparison of reference variables for forecasting one-year ahead sales growth

Reference Variable	Window	Size	Δ_q (rank)	KS (rank)	CvM (rank)
opmarΔ_6	30	0.025	0.0155 (1)	1.874 (4)	0.8265 (3)
opmarΔ_7	30	0.025	0.0157 (2)	2.1986 (10)	1.0815 (8)
opmarΔ_6	30	0.01	0.0161 (3)	2.4469 (14)	1.3149 (13)
operating margin	10	0.05	0.0279 (24)	4.1606 (50)	6.1461 (74)
operating margin	5	0.05	0.0303 (26)	4.4720 (74)	4.8603 (43)
sales (Mauboussin)	5	–	0.0516 (125)	6.3825 (213)	12.7518 (199)
sales	5	0.05	0.0524 (133)	6.3939 (214)	13.4453 (212)
major group	5	–	0.0653 (201)	8.6576 (274)	22.5482 (256)
industry group	5	–	0.0935 (295)	10.7868 (302)	36.6514 (291)

Table 2.6.: Comparison of reference variables for forecasting three-year ahead sales growth

Reference Variable	Window	Size	Δ_q (rank)	KS (rank)	CvM (rank)
opmarΔ_7	30	0.025	0.0286 (1)	3.2227 (8)	2.8868 (10)
opmarΔ_8	30	0.025	0.0301 (2)	1.9903 (2)	1.0989 (1)
opmarΔ_8	30	0.01	0.0302 (3)	1.9878 (1)	1.2532 (4)
operating margin	30	0.01	0.0598 (29)	6.9177 (65)	16.7632 (63)
operating margin	5	0.05	0.0697 (38)	10.4675 (160)	33.6971 (119)
operating margin	10	0.05	0.0877 (73)	11.8366 (200)	55.6297 (200)
sales (Mauboussin)	5	–	0.1028 (143)	13.4856 (247)	61.3185 (211)
sales	5	0.05	0.1057 (155)	13.8816 (253)	63.7592 (213)
major group	5	–	0.1423 (274)	17.9423 (311)	106.9768 (292)
industry group	30	–	0.1863 (309)	16.9141 (302)	117.9496 (302)

Table 2.7.: Comparison of reference variables for forecasting five-year ahead sales growth

Reference Variable	Window	Size	Δ_q (rank)	KS (rank)	CvM (rank)
opmarΔ_{10}	30	0.01	0.0312 (1)	2.204 (3)	1.3081 (2)
opmarΔ_{10}	30	0.025	0.0341 (2)	1.7507 (1)	0.9922 (1)
opmarΔ_6	30	0.01	0.0361 (3)	2.4614 (6)	2.0039 (9)
operating margin	30	0.01	0.0851 (37)	9.4868 (89)	32.0685 (84)
P/E	5	0.05	0.1096 (55)	9.2194 (88)	41.3370 (93)
P/E	10	0.025	0.1485 (128)	12.5293 (133)	79.8237 (152)
sales (Mauboussin)	5	–	0.1600 (170)	19.0380 (277)	137.3941 (261)
sales	5	0.05	0.1650 (187)	19.5103 (279)	147.1779 (269)
major group	30	–	0.2136 (289)	16.7058 (243)	106.9918 (231)
industry group	30	–	0.2179 (296)	17.6483 (261)	127.3253 (255)

Table 2.8.: Comparison of reference variables for forecasting 10-year ahead sales growth

Reference Variable	Window	Size	Δ_q (rank)	KS (rank)	CvM (rank)
opmarΔ_6	30	0.025	0.0432 (1)	3.7904 (5)	4.1498 (5)
opmarΔ_7	30	0.025	0.0456 (2)	3.5849 (3)	3.8386 (2)
opmarΔ_5	30	0.025	0.0478 (3)	4.0971 (15)	5.0842 (9)
operating margin	30	0.01	0.1112 (36)	7.4423 (80)	20.6308 (88)
opmarΔ_{10}	10	0.025	0.2033 (113)	8.5930 (103)	31.8499 (106)
sales	30	0.01	0.2099 (115)	10.0584 (112)	42.9767 (118)
sales (Mauboussin)	30	–	0.2270 (128)	11.2416 (130)	50.6546 (128)
major group	30	–	0.2561 (146)	12.0198 (131)	61.4773 (134)
P/E	5	0.01	0.2842 (168)	17.4874 (183)	136.6147 (192)
industry group	30	–	0.2859 (169)	13.4787 (141)	75.4007 (145)

The approach by Mauboussin and Callahan (2015) performs in a very similar way to using *sales* as a reference variable in our approach. For forecast horizons one, three and five their approach is slightly better than ours using *sales* and for a 10-year horizon it is vice versa. Nonetheless, their approach performs clearly worse than the best parameter combinations according to our accuracy measures.

Although it is not the aim of this work to give a theoretical framework of the drivers of sales growth, we try to give some intuition behind the results presented above, especially as the operating margin or its past delta are not commonly known as drivers of sales growth. Both figures are cumulative metrics which condense a lot of information, for example, the competition within the industry (see, e.g., Porter, 1979) or the competitive position of the company (see, e.g., Porter, 1985) which both significantly affect the operating margin (deltas) as well as the future development of a company. Intuitively, the more a company's operating margin grows the better its market position is and it is natural to expect a higher sales growth. This corresponds to the results in Table 2.9 discussed below. Thus, it is not too surprising that the reference variables *operating margin* and *past operating margin deltas* perform better than other variables including much less information. With respect to the benchmark approach of Mauboussin and Callahan (2015) the superior performance could be partly explained by Gibrat's law which basically states that the proportional rate of growth of a company is independent of the absolute size (Gibrat, 1931).

To get a feeling for the influence of the reference variable in our new approach on the shape of the distribution forecast provided by the reference class, we consider the year 2018 as an example in view of the later application in practice. For each forecast horizon

Table 2.9.: Influence of the best reference variables on median, mean and standard deviation of the reference classes for forecasting compound sales growth for different forecasting horizons. Mean and standard deviation are 2.5% trimmed on both tails. opmarΔ_7 is measured in percentage points per year in this table.

	one-year forecast horizon				three-year forecast horizon			
qu.	opmarΔ_6	median	mean	std	opmarΔ_7	median	mean	std
10%	-3.50	-0.04	1.65	26.72	-2.74	0.43	0.43	17.66
20%	-1.44	0.66	1.28	17.58	-1.19	0.81	0.86	11.83
30%	-0.74	1.39	1.97	14.62	-0.62	1.68	2.12	10.17
40%	-0.33	2.39	3.04	13.98	-0.28	1.93	2.20	10.03
50%	-0.03	3.40	4.64	12.47	-0.02	2.55	3.06	9.57
60%	0.27	3.16	4.18	12.62	0.23	2.48	3.01	9.43
70%	0.68	3.77	4.93	14.07	0.58	2.92	3.73	9.73
80%	1.44	3.66	5.23	17.69	1.20	3.34	4.34	12.28
90%	4.48	4.67	7.36	28.57	3.51	4.16	5.31	17.88

	five-year forecast horizon				10-year forecast horizon			
qu.	opmarΔ_{10}	median	mean	std	opmarΔ_6	median	mean	std
10%	-1.74	0.40	0.59	11.84	-2.68	1.49	1.32	10.82
20%	-0.83	1.27	1.20	9.58	-1.19	2.37	2.68	6.75
30%	-0.47	2.31	2.48	8.55	-0.63	1.58	1.78	6.36
40%	-0.22	1.44	1.56	8.57	-0.27	2.46	2.68	6.25
50%	-0.04	2.04	2.15	7.14	0.00	2.73	2.59	5.96
60%	0.14	3.24	3.40	8.44	0.27	3.02	3.42	6.16
70%	0.37	1.87	2.49	7.96	0.63	2.96	3.28	6.49
80%	0.75	2.69	3.21	8.81	1.27	3.04	3.58	7.19
90%	2.05	3.15	4.55	13.36	3.59	4.82	4.97	10.55

we use the best parameter combination, according to the measure of quantile deviations Δ_q and construct artificial initial cases by calculating the 10% to 90% quantiles of the reference variable. After that, we use our new approach to construct reference classes based on these initial cases. Table 2.9 displays the value of the reference variables and the median, mean and standard deviation of the distributional forecast of the associated quantiles.

The location and scale parameters behave similarly for all forecasting horizons. The standard deviation is smallest for medial reference variables and rises towards the tails reflecting the uncertainty in the tails of the distributions by this v-shape. The mean and median are monotone in the reference variable quantiles besides few exceptions indicating that higher past margin deltas coincide with higher sales growth.

2.5. The Outside View in Practice

In the last section we systematically investigated the accuracy of constructing reference classes using a single reference variable. In practice, we are able to assess a prediction by evaluating the empirical distribution function of the reference class. Thus, we can use the distributional information, that is, the outside view, of the reference class to correct a potentially flawed or biased prediction. Moreover, we can calculate point forecasts based on the median or mean of the reference class, confidence intervals based on the quantiles of the distributional forecast, or similarity-based forecasts using the outcomes of the reference class by weighting them according to a measure of similarity to the initial case.

However, in order to demonstrate how to use our method in practice, we compare the resulting outside view with experts' forecasts and calculate base rates for two examples – 3M and Amazon. To be more precise, for both companies we forecast the distribution of one-year annual sales growth based on the best combination of reference variable and hyper parameters. These results are compared to analysts' forecasts which were obtained from the FactSet (2021) estimates database, whereby for both estimates 2018 is the base year.[5] The results are presented in Figures 2.6 and 2.7.

For 3M there are 15 expert forecasts and Figure 2.6 illustrates that these forecasts vary between -2.35% and 3.26% and lie slightly below the median of our forecasted distribution. Thus, there is no indication of overoptimistic forecasts as in- and outside views coincide. Both views classify 3M as an average company with respect to sales growth. However, the low variability of forecasts may lead investors to overconfidence in the reported range of forecasts. The outside view uncovers higher sales growth variability, thus preventing the overconfidence pitfall.

Figure 2.7 shows the results for Amazon, based on 43 expert forecasts, which differ considerably. On the one hand, the forecasts are more heterogeneous and vary between 13.93% and 22.82%. On the other hand, the forecasts are much more optimistic and correspond to quantiles between 76.87% and 88.25%. This means that for the most

[5] We also calculated the distribution for the three-year sales growth but the results are very similar with respect to the basic statement, thus we only report the one-year results. Moreover, we could not take longer prediction horizons into account as there were far too few observations available.

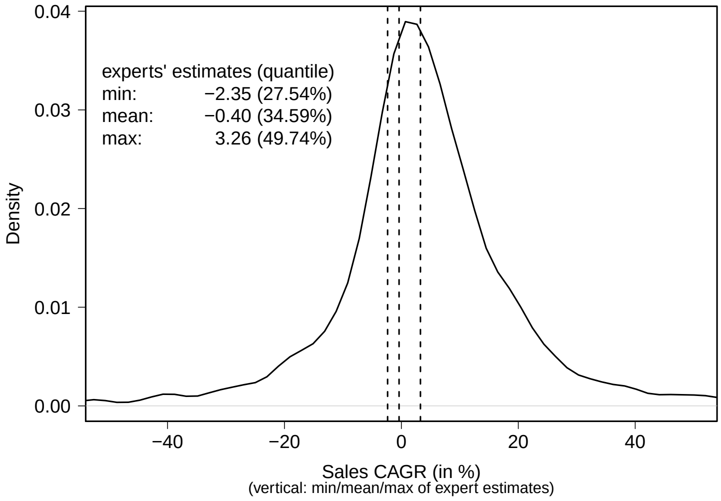

Figure 2.6.: Forecasted density of one-year sales growth for 3M based on six-year operating margin delta (1.77 percentage points) and with hyper parameters window = 30 and size = 0.025 compared to experts' estimates. For density estimation on support $[-100, \infty)$ we used the Gaussian kernel with Silverman's rule of thumb as bandwidth.

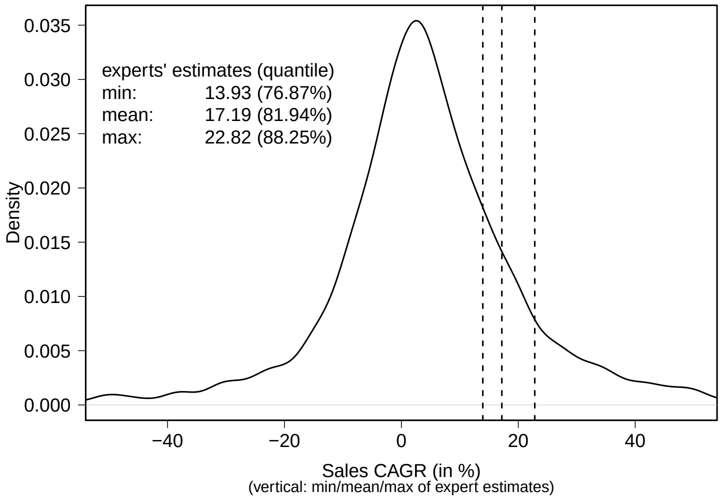

Figure 2.7.: Forecasted density of one-year sales growth for Amazon based on six-year operating margin delta (4.16 percentage points) and with hyper parameters window = 30 and size = 0.025 compared to experts' estimates. For density estimation on support $[-100, \infty)$ we used the Gaussian kernel with Silverman's rule of thumb as bandwidth.

Table 2.10.: Comparison of base rates for 3M based on reference classes of our approach using the respective best reference variable and hyper parameters, and of Mauboussin and Callahan (2015). Mean and standard deviation are 2.5% trimmed on both tails.

3M	Base Rates							
CAGR (%)	1-Yr	1-Yr MC	3-Yr	3-Yr MC	5-Yr	5-Yr MC	10-Yr	10-Yr MC
≤ -25	4.13	4.64	2.12	1.53	1.16	0.97	0.57	0.41
]-25,-20]	1.50	1.71	1.77	2.39	0.66	0.83	0.43	0.26
]-20,-15]	2.71	2.92	1.58	4.11	2.64	2.07	1.42	1.31
]-15,-10]	4.01	4.42	3.89	5.40	3.31	4.77	2.83	2.90
]-10,-5]	7.86	8.72	8.87	10.67	8.43	11.20	6.02	9.65
]-5,0]	16.16	19.37	18.04	26.13	18.51	27.37	17.08	27.77
]0,5]	20.17	24.17	24.25	26.07	32.23	29.72	35.79	35.65
]5,10]	14.95	15.95	15.57	13.62	15.70	15.41	20.55	15.72
]10,15]	9.96	6.46	9.41	5.09	8.43	3.87	8.79	4.11
]15,20]	6.36	3.48	5.47	2.21	4.63	2.07	3.97	1.32
]20,25]	3.73	2.48	3.65	1.10	2.64	0.76	1.63	0.51
]25,30]	2.15	1.55	1.53	0.67	0.66	0.41	0.57	0.27
]30,35]	1.58	1.16	1.43	0.18	0.99	0.35	0.14	0.09
]35,40]	1.05	0.77	0.59	0.18	0.00	0.14	0.14	0.02
]40,45]	0.45	0.72	0.69	0.12	0.00	0.00	0.07	0.00
> 45	3.24	1.49	1.13	0.49	0.00	0.07	0.00	0.00
mean	4.30	1.59	3.54	-0.45	2.57	0.29	3.18	0.89
median	3.33	1.73	2.53	-0.04	2.38	0.31	3.02	0.92
std	12.89	11.31	10.09	7.80	7.66	6.32	6.30	5.13

optimistic forecast, roughly only one out of 10 companies within the reference class managed to reach the forecasted growth of Amazon. This big difference between in- and outside views should at least exhort the analysts to scrutinize their forecasts and to question the arguments for the optimistic assessment. Although Amazon is well known to be a high-growth company the analysts should have good reasons for such optimistic forecasts.

Tables 2.10 and 2.11 are inspired by Mauboussin and Callahan (2015) and show the base rates for 3M and Amazon. At this point it is worthwhile mentioning that our method yields different base rates for each company while the method of Mauboussin and Callahan results only in 11 clusters with one set of base rates for each. Furthermore, it is noteworthy that for both companies, and every forecast horizon, the mean, median

Table 2.11.: Comparison of base rates for Amazon based on reference classes of our approach using the respective best reference variable and hyper parameters, and of Mauboussin and Callahan (2015). Mean and standard deviation are 2.5% trimmed on both tails.

Amazon	Base Rates							
CAGR (%)	1-Yr	1-Yr MC	3-Yr	3-Yr MC	5-Yr	5-Yr MC	10-Yr	10-Yr MC
≤ -25	4.37	3.31	1.72	1.23	1.16	2.08	1.06	0.35
]-25,-20]	1.74	0.55	1.77	3.68	0.83	0.00	0.57	0.71
]-20,-15]	2.39	3.87	1.72	4.29	2.31	4.17	1.63	2.36
]-15,-10]	4.50	2.76	3.55	4.29	2.81	3.47	2.20	2.60
]-10,-5]	8.95	8.29	7.93	11.04	8.60	15.97	6.87	10.64
]-5,0]	14.54	17.68	19.02	19.02	18.35	16.67	20.84	30.02
]0,5]	18.47	26.52	22.77	28.22	33.06	32.64	32.67	33.33
]5,10]	13.69	16.02	18.43	17.79	15.87	19.44	20.77	16.31
]10,15]	10.04	6.63	9.96	5.52	9.09	4.17	6.52	2.84
]15,20]	6.97	4.42	5.32	3.07	2.98	0.00	4.46	0.71
]20,25]	3.93	5.52	2.37	1.23	2.31	0.69	1.35	0.12
]25,30]	2.59	1.66	1.38	0.61	1.32	0.69	0.50	0.00
]30,35]	1.94	1.10	1.28	0.00	0.50	0.00	0.50	0.00
]35,40]	1.26	1.10	0.84	0.00	0.50	0.00	0.00	0.00
]40,45]	1.09	0.55	0.34	0.00	0.17	0.00	0.00	0.00
> 45	3.52	0.00	1.58	0.00	0.17	0.00	0.07	0.00
mean	4.93	2.75	3.72	0.00	2.55	0.03	2.65	0.16
median	3.70	2.27	2.88	0.39	2.16	1.23	2.50	0.49
std	14.11	10.73	9.74	8.23	7.62	7.01	6.46	5.15

as well as standard deviation are higher for our reference classes. This is due to the fact that small firms are included within our reference classes. This observation is in line with the results presented in Mauboussin and Callahan (2015) where these figures also increase with decreasing sizes of companies. As 3M and Amazon are relatively large companies with sales of 32.7 billion USD and 232.9 billion USD in 2018, respectively, small companies are not included in the reference classes of Mauboussin and Callahan. As a further consequence, the base rates of our approach are less concentrated in the range -5% to 10% and imply a wider range of possible outcomes which appears realistic. However, we do not want to make an assessment of the procedures as this point as this was already done within the last section.

2.6. Conclusion and Outlook

In this chapter, we have extended financial analysts and investors' toolbox by a general method to provide outside views for forecasting sales growth and we have provided an extensive backtest on sales data from the USA over several decades. Additionally, we have compared the method to several benchmark approaches used in practice and applied it to real world examples of 3M and Amazon. The new approach delivers very reasonable results, needs only a parsimonious amount of data and is easy to interpret. Thus, it is well suited to applications in practice and lays a sound foundation for further research as several extensions of our approach are possible.

First, the method itself can be extended by including multiple reference variables or time series characteristics. In our approach, we focus on the case of one variable having an easy interpretation and a direct extension of the approach by Mauboussin and Callahan (2015) in mind. Clearly, it would be interesting to see if better reference classes could be constructed with more than one reference variable.

Within our method, the crucial part is to find orderings of the forecast ability of the different reference variables based on several quality criteria. We have not answered the question in which sense the different forecasts are statistically significantly different. Moreover, it is still an open question which reference variables are actually acceptable for generating appropriate outside views and which not. That means, it would be interesting to know in which numerical regions the goodness-of-fit measures may or may not lie, especially for the new ranking based on Δ_{q}. Perhaps, a testing approach for relevant differences like Dette and Wied (2016) could be helpful here. The thresholds could be determined by potential losses induced by correcting the experts' forecasts which Kahneman and Tversky (1979) propose.

Finally, several stress tests of our method are possible. One could perform a simulation study to assess how well reference classes can uncover true underlying distributions of any variable in order to better understand the mechanics of reference classes. Furthermore, a formal approach of correcting potentially biased expert forecasts with the similarity-based outside views can be worked out and backtested. This means that one would consider point forecasts based on the median or mean of the forecasted distributions, combine them suitably with the experts' views and backtest whether these combinations lead to better overall forecasts.

Chapter 3.

Distributional Reference Class Forecasting of Corporate Sales Growth With Multiple Reference Variables

3.1. Introduction

A major aspect of statistics is to make projections and forecasts of future events which should be probabilistic in nature to reduce uncertainty (Dawid, 1984). To this end, we extend the method for distributional forecasts with reference classes proposed in Chapter 2 to allow for reference class construction based on several co-variates. In this context, a reference class supplies the outside view, that is, statistical information or empirical data on similar past forecast challenges where the outcomes have been observed (Kahneman and Tversky, 1979). The outside view then provides a distributional forecast based on the distribution of outcomes within the reference class. Additionally, the resulting reference classes may be used to make interval predictions (c.f. Zarikas and Kitsos, 2015) or point forecasts (e.g., using the framework by Gilboa et al., 2006). In contrast to reference class forecasting, a forecast based on the inside view concentrates on the uniqueness and specificity of the given forecast challenge (Kahneman and Lovallo, 1993).

In forecasting practice, the use of base rates, that is, distributional information, is recommended (Armstrong, 2005) and reference classes and outside views are known in the literature; for example, Bordley (2014) uses distributional information from a reference class as a Bayesian prior in the context of healthcare cost. Chapter 2 reviews the literature on reference class selection, quintessentially, there is a lack of studies that systematically investigate the objective choice of reference classes. Such investigation is necessary from an optimality point of view and, further, because Kent et al. (2018) discuss risks arising from ill-advised reference class selection in the context of randomized clinical trials. In this chapter, we revisit forecasting corporate sales growth to allow for a direct comparison of new results to the findings in Chapter 2. Sales growth rate predictions, often by experts with an inside view, suffer from low predictability and overoptimism (Chan et al., 2003). Possible further applications of reference class forecasting in finance include corporate bankruptcy prediction (Hull, 2020), stock returns (Bordalo et al., 2019) or cash flow items (Guerard et al., 2015) that are pivotal features in fundamental analysis of firms, key drivers of stock selection models and in general challenging forecasting tasks where long forecast horizons increase the level of complexity.

In general, constructing a reference class of comparable objects is known as the reference class problem in statistics. While forecasting probabilities with respect to a given object, Venn (1888) notes that each object has several characteristics to determine a set of similar objects from which to derive these probabilities. Reichenbach (1949) first called such a set a reference class. Here, we consider firms and can imagine plenty of possible reference classes, e.g., all US firms, all S&P500 firms, or firms with similar cash flow or stock market metrics. Thus, we are challenged to find a reference class that is best in the sense of forecasting the distribution of, e.g., three-year sales growth.[1] Clearly, the search for a reference class is paired with a specific forecast challenge in mind and a *good* reference class always depends on this forecast challenge.

We propose a framework for reference class selection based on multiple co-variates, here called reference variables, and examine several approaches to find outside views to forecast corporate sales growth. Hence, we construct reference classes of additional observations that share similarities to the company at hand with respect to multiple reference variables based on new rank-based algorithms that allow for an optional preprocessing data dimension reduction through principal component analysis. These approaches are easy to implement and we choose interpretable algorithms to build the reference classes. Thus, the proposed methods are well suited for practical application, the more so as the outside view is straightforwardly provided by the realized sales growth rates within the reference class. In line with Chapter 2, we recommend distributional forecasts in terms of the empirical cumulative distribution function (ECDF) of the reference class outcomes to reduce uncertainty. ECDFs are easy to calculate, non-parametric and include no assumptions on the underlying distribution. Their simple structure empowers practitioners to investigate the reference class for a given company and discuss the nature of forecasts highlighting the procedure's interpretability. Alternatively, a parametric model for sales growth distributions is discussed in Stanley et al. (1996).

[1] Venn (1888) originally describes an example of a fifty-year-old consumptive Englishman with many possible reference classes, for example, all humans, all males, all at least fifty-year-old Englishmen or all consumptive patients, that could be used for a distributional forecast of, e.g., remaining life expectancy.

The forecast performance of the new rank-based algorithms using different sets of reference variables is backtested on the same data set as in Chapter 2 (see Section 2.3) and calibration is ranked by Δ_{q} as well. A review of distributional forecast evaluation places Δ_{q}, a measure based on probability integral transform values, within the context of the literature thereon. In the backtest, especially past operating margins and past sales growth rates are the best performing variables for reference class building with respect to future sales growth rates and a subsequent distributional forecast thereof. Dimension reduction using principal component analysis allows for using more variables, for example, contemporaneous balance sheet and financial market parameters, and shorter lags of past variables while simultaneously improving the results substantially by between 38% and 71%, depending on the forecast horizon. Further, an application of the new algorithms on sales growth demonstrates distributional reference class forecasting in practice, shows how to forecast intervals and compares distributional forecasts to analysts' estimates, thus, illustrating the utility of reference classes and how to apply their results in practice. We additionally illustrate the retrospective use of historic distributional forecasts of sales growth rates in comparison to realized values.

This chapter is organized as follows: Section 3.2 contains the theoretical framework of reference class selection with multiple reference variables and the newly proposed algorithms. Section 3.3 reviews relevant literature on performance measures for distributional forecasts. Further, Section 3.4 presents the backtest along with the variable and model selection procedure. Illustrative practical applications are demonstrated in Section 3.5. Finally, Section 3.6 concludes and gives an outlook on future research.

3.2. Distributional Reference Class Forecasting

The aim of reference class forecasting as proposed by Kahneman and Tversky (1979) is to obtain well-behaved predictions with respect to an initial forecast case that are based on a reference class. For this purpose, a reference class consists of past forecast cases similar to the inital case which need to provide statistical information on the variable to be predicted, i.e., the outside view. Then, the reference class forecast may be used to assess expert or model-based forecasts and to potentially correct them (see Tetlock and Gardner, 2016). Expert forecasts often take an inside view, based on the specific characteristics of the forecast case. Thus, reference class forecasting provides a data driven method to overcome negative impacts of the inside view such as overoptimism, wishful thinking or strategic misrepresentations (see, e.g., Tversky and Kahneman, 1974;

Kahneman and Tversky, 1973). Intending to reduce uncertainty, we prefer to use reference classes for distributional forecasts.

In literature, there are several studies dealing with reference class forecasting; e.g., Batselier and Vanhoucke (2016), Servranckx et al. (2021) and Natarajan (2022) select reference classes in project planning or Lovallo et al. (2012) for stock returns where all selections are based on data availability and at most subjective categories (see Chapter 2 for an extended literature review). Notably, Knudsen et al. (2017) discuss an objective choice of peer groups in corporate valuation based on similarity but choose only six firms as peers which makes the reference classes prone to bias. Chapter 2 provides a systematic analysis of reference class forecasting by backtesting different but only a single reference variable. In view of the reference class problem (Venn, 1888; Cheng, 2009, for a more recent review) and in strive for optimal reference class forecasts it is self-evident that the considerations should not be limited to similarity with respect to a single co-variate of the initial case. Here, we extend the existing framework from Chapter 2 to select appropriate reference classes for distributional forecasts to the case of multiple reference variables. However, a careful assessment thereof is necessary. If reference classes consist only of elements extraordinarily similar to the initial object, there is a risk of undersized and little informative reference classes producing likewise biased forecasts.

For a multiple variable reference class selection, we propose different rank-based algorithms that allow for using multiple reference variables including an optional dimension reduction. The approaches are easy to implement and interpret and find reference classes for each examined company separately. Then, an assessment of the distribution within the reference classes follows directly from the outcomes within the reference class in shape of their ECDF that serves as a distributional forecast. In order to demonstrate the advantages of the new rank deviation procedures, we consider forecasting sales growth and compare the new approaches to the method from Chapter 2. To this end, we identify optimal combinations of reference variables and algorithm options for reference class selection and evaluate them within a backtest on the same data set as described in Section 2.3 for a meaningful quality comparison. Table 2.1 describes all potential reference variables. Note that Table 2.2 displays compound annual sales growth rates and annual means of operating margin deltas but we use non-averaged sales growth rates and operating margin deltas for reference class selection. The following two subsections extend the theoretical foundation of Chapter 2 and propose algorithms to select reference classes.

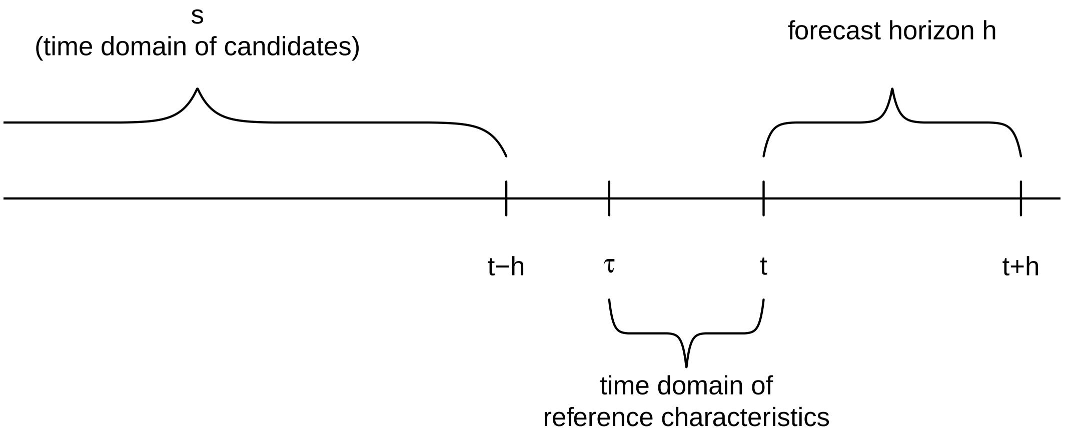

Figure 3.1.: Illustration of reference class and prediction timeline. Firms j at times s denote the set of potential members (candidates) for the reference class of individual i at time t. Note, $\tau \leq t-h$ is possible as well if $\tau \geq h$.

3.2.1. Theoretical Framework for Multiple Reference Variables

The framework given in Section 2.2.1 extends naturally to the case of multiple reference variables. For a firm i at time t, we construct a reference class to generate a forecast of the distribution of $Y_{i,t+h}$, that is a distributional h-step ahead forecast of the random variable $Y_{i,t}$, here sales growth, for individual i at time t.[2] To this end, we assume that a sufficient amount of historical data on firms is available to assess the distribution of $Y_{i,t+h}$. We base the reference class on reference variables $X_{i,\tau:t} := \{X_{i,t'}\}_{t'=\tau,\dots,t}$ and build a reference class R by finding firms j in the past which are similar to individual i at time t with respect to the reference variables. Similarity can be measured in multiple ways, for mathematical purposes it is convenient to view similarity according to some distance measure $d : D^2 \to [0,\infty)$. Then, $d(X_{i,\tau:t}, X_{j,\zeta:s})$ shall be *small*, where $s+h \leq t$ ensures that the realization of $Y_{j,s+h}$ is available and D is the domain of $X_{i,\tau:t}$ (c.f. Figure 3.1). d could be a metric, e.g., based on some norm. A non-parametric forecast for the distribution of $Y_{i,t+h}$ is now given by the empirical cumulative distribution function of the values $Y_{j,s+h}$, $(j,s) \in R$ and serves as an outside view.

Reference class forecacsting requires the assumptions of Chapter 2 regarding the depen-

[2] We phrase the theoretical framework with a specific application in mind, namely forecasting corporate sales growth. For a general purpose the term ‘firm’ can be replaced by ‘object’ and the term ‘sales growth’ can be replaced by ‘predictand’.

dence of $Y_{i,t}$ on $X_{i,\tau:t}$ and the stability of this dependence over time. These are a data generating mechanism[3] modeled by a continuous function f_h such that $Y_{i,t+h} \sim f_h(X_{i,\tau:t})$ and a stationarity assumption such that $Y_{j,s+h} \sim f_h(X_{j,\zeta:s})$, $(j, s) \in R$, within the reference class. If $d(X_{i,\tau:t}, X_{j,\zeta:s})$ is small, then $f_h(X_{i,\tau:t})$ is close to $f_h(X_{j,\zeta:s})$ and the empirical distribution of $Y_{j,s+h}$, $(j, s) \in R$, is a good approximation for the distribution of $Y_{i,t+h}$. Moreover, $f_h(X_{i,\tau:t})$ can be interpreted as the conditional distribution of $Y_{i,t}$ given $X_{i,\tau:t}$ and stationarity ensures a stable data generating process over time.

3.2.2. Proposed Algorithms for Reference Class Selection

Algorithms for constructing a reference class from a given sample need to implement the aforementioned assumption regarding the stable dependence of sales growth $Y_{i,t}$ on reference variables $X_{i,\tau:t}$ and need to decide which past firm observations are similar enough with respect to the reference variables. A window length parameter w common to all algorithms defines the number of past years to use for a specific forecast challenge. w selects observations from the sample to constitute a set of candidates C for the reference class from a limited time period and thereby accounts for the degree of stability regarding the dependence. Assessing the similarity to the firm of interest and deciding whether it is part of the reference class or not is the essential feature of each algorithm.

The decision problem of labeling each candidate 'belonging to reference class' and 'not belonging to reference class' makes the reference class selection a binary classification. The selection is based on available co-variates (reference variables) only and not on the outcome of the candidate firms because the outcome is naturally not observed for the initial firm at hand. Thus, we use unsupervised learning techniques to find firms that are sufficiently similar to the initial firm by algorithms. The decision on sufficient similarity is part of the reference class problem as it chooses a subset of candidates to form the reference class.

In the application on firms we encounter skewed reference variables including outliers (see Table 2.2)[4] and use rank-based methods to be robust against these data features. Unsupervised cluster algorithms share the property to split the set of candidates in a

[3] In case of a finance application like here, such a data generating mechanism may be called *market mechanism*.

[4] In addition to Table 2.2, negative skew occurs for the reference variables operating margin, β, one- and two-year operating margin delta and price-to-earnings ratio which has the smallest skew with roughly -167. All other reference variables have a positive skew with up to roughly 400 in the case of price-to-book ratio. This supports the use of rank-based methods to reduce skewness effects.

fixed number of clusters and due to continuity of reference variables in our case we argue that it does not necessarily make sense to find candidates that are closest to the given firm in this manner (c.f. Figure 2.1)[5].

We proceed with the following definition that a reference class has to consist of at least 20 elements or members in order to make reasonable distributional forecasts.

Definition 1. *Reference Class*
Let C be a set of reference class candidates. We call any set $R \subset C$ of observations j at times s reference class if the reference variables $X_{j,\zeta:s} = \{X_{j,s'}\}_{s'=\zeta,\dots,s}$ and the outcomes $Y_{j,s+h}$ are observed and $|R| \geq 20$.

The set of candidates for a reference class is the largest possible reference class (in sense of cardinality). It includes all objects that could potentially be a member of the reference class and additionally serves as a market climate reference class that captures the overall market sentiment for the time period of candidate firms. The resulting ECDF may serve as an estimate of the marginal distribution neglecting any confounding variables.

The new rank deviation procedures using multiple reference variables take the backtest against the market climate reference class, the method presented in Chapter 2 for single reference variables, the approach of Mauboussin and Callahan (2015) basing similarity on fixed sets of sales levels (see Section 2.2) and the group approach (see Section 2.4) using the major or industry group of a firm to select all candidate firms for the reference class that are in the same respective group as the initial firm. Thus, membership in the same major or industry group is said to fulfill the assumption of sufficient similarity. The approach is in line with common practice in corporate valuation to form peer groups based on industry classification due to the assumption that firms in the same industry are similar in terms of value determinants (Bhojraj and Lee, 2002; Marozzi, 2011).

3.2.2.1. Rank Deviations

We introduce a novel method using rank deviations that assesses similarity of candidate firms based on multiple reference variables and extends the approach in Section 2.2.1

[5] This applies to the special case here. The proposed procedures do not account for ‘natural’ clusters that might occur. For example, there might be only 20 observations ‘very similar’ to the initial case. But the algorithm may choose the most similar 25 observations and, thus, five less informative observations. Hence, there is no general conclusion and the algorithms must be adapted to the specific forecast challenge.

which uses an arbitrary but single reference variable. Time series data in discrete time can be incorporated by treating each point in time as an additional reference variable. The novel method is rank-based to mitigate skewness effects as well as outlier influence on the selection and constructs custom reference classes seperately for each forecast challenge, i.e., each inital firm here.

Sufficient similarity is measured based on ranks and a size parameter $c \in (0, 1)$ that controls the size of the reference class as a fraction of the candidate set exploiting the continuity of reference variables. Consequently, the parameter c determines which of the candidates' reference variables $X_{j,\zeta:s}$ lie closely enough to the initial firm's reference variables $X_{i,\tau:t}$ to be a member of the reference class. Thus, c assesses for which candidate firms j at time s the value $d(X_{i,\tau:t}, X_{j,\zeta:s})$ is considered as *small*. The case of a single reference variable illustrates the method. We select a fraction c of candidates $(j, s) \in C$ for the reference class with the least absolute rank deviation $|R(X_{i,t}) - R(X_{j,s})|$ as sufficiently similar to the initial firm. Here, the rank function $R : \mathbb{R} \to [1, |C| + 1]$ calculates the rank of a single reference variable in the set of candidate firms and the initial firm. This is equivalent to the procedure in Section 2.4 where candidate firms are selected by a single variable and the reference class consists of the fraction c of candidates closest to the initial firm's observation with respect to the empirical quantile function of all candidate firms.

We propose three ways of combining $\kappa > 1$ reference variables by intersecting or unifying the reference classes obtained from several single reference variables or by using the candidate firms that have least absolute rank deviation (LARD) inspired by Knudsen et al. (2017). The two set-theoretic operations both involve first constructing κ reference classes based on each reference variable seperately. On the one hand, we combine the reference classes by intersecting them with the possibility of having few or none observations left. Constructing the initial reference classes with an adjusted $c_{\text{inter}} = \min\{c\kappa, 0.25\}$ may avoid an insufficient amount of remaining observations. On the other hand, we combine the reference classes by union where selecting too many candidates may be solved by constructing the initial reference classes with an adjusted $c_{\text{union}} = c/\kappa$. Unifying the reference classes has the additional advantage that not all reference variables must be observed for each reference class candidate. The application of LARD requires a ranking of candidate firms and inital firms according to each reference variable resulting in rank vectors $r_{i,t} = \mathcal{R}(X_{i,\tau:t})$ for the initial case and $r_{j,s} = \mathcal{R}(X_{j,\zeta:s})$ for all candidates $(j, s) \in C$, where $\mathcal{R}$ is the rank function applied on each entry of the reference variables seperately. Then, the fraction c of observations with the least absolute rank deviation

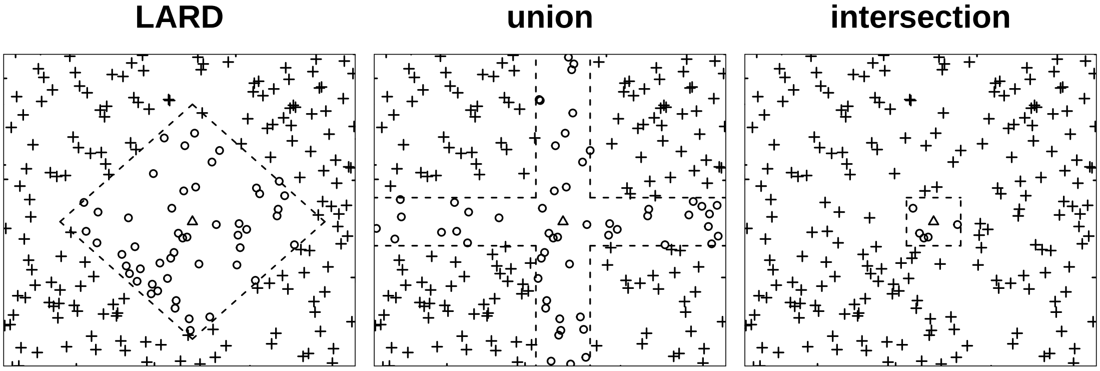

Figure 3.2.: Illustration of the three rank deviation methods for reference class selection based on $\kappa = 2$ reference variables on them same candidate set with size $N = 500$ and reference class size $c = 0.1$. Ranks of the reference variables between 100 and 400 are displayed on the horizontal and vertical axes. The triangle shows the initial firm, circles are selected as reference class members and crosses are the remaining observations.

$d_{j,s} = |r_{j,s} - r_{i,t}|$ in L_1 norm $||d_{j,s}||_1$ constitutes the reference class. The algorithm is related to the k nearest neighbors algorithm where $k = cN$ is chosen relative to the number of candidates N and proximity is measured by L_1-norm of ranks without subsequent regression or classification but with a distribution forecast. Other norms could be used but L_1 is a natural choice in combination with ranks. Further, intersecting κ reference classes is related to the supremum norm of the vectors of absolute rank deviations $d_{j,s}$ and the union of κ reference classes is comparable to selecting reference classes by the minimum entries of $d_{j,s}$ (c.f. Figure 3.2 for both).

3.2.2.2. Principal Component Analysis Rank Deviation

In order to use information from a large number of co-variates, we first apply principal component analysis (PCA) to reduce the dimensionality of the problem and then use the rank deviation procedures on the rotated data to identify the reference class. Combing several reference variables by LARD is related to the k nearest neighbors algorithm which is used in algorithmic pipelines with PCA, e.g., in facial recognition (Marcialis and Roli, 2004; Parveen and Thuraisingham, 2006). Although the union procedure allows for different sets of reference class candidates for each co-variate, this is no longer the case for PCA preprocessing. Constructing the reference class based on principal components (PCs), all variables must be available in order to rotate the original data matrix.

PCA is carried out on the original reference variables, a transformed set of reference variables $\mathcal{X}_{j,\zeta:s}$ or a subset C' of the candidate set C obtained by one of the following four initial transformations: a) no inital transformation, that is, $\mathcal{X}_{j,\zeta:s} = X_{j,\zeta:s}$ and $\mathcal{X}_{i,\zeta:t} = X_{i,\tau:t}$; b) use the initial transformation $\mathcal{X}_{j,\zeta:s} = X_{j,\zeta:s}^{1/5}$ and $\mathcal{X}_{i,\tau:t} = X_{i,\zeta:t}^{1/5}$ to mitigate skewness effects in the data; c) compute ranks $\mathcal{X}_{j,\zeta:s} = \mathcal{R}(X_{j,\zeta:s})$ and $\mathcal{X}_{i,\tau:t} = \mathcal{R}(X_{i,\tau:t})$ for each reference variable seperately; or d) trim the data for each reference variable across the candidate set by 2.5% on each tail and then reduce the candidate set C by all candidates that get trimmed in at least one reference variable such that all observations are complete in the subset of remaining candidates $C' \subset C, |C'| = N'$.

Let $\mathcal{X}$ be the $N \times \kappa$ data matrix containing the potentially transformed reference variables from the set of (remaining) candidates C, and let W be the $\kappa \times \kappa$ weight matrix whose columns are the eigenvectors of the correlation matrix $(N-1)^{-1}\tilde{\mathcal{X}}'\tilde{\mathcal{X}}$ of $\mathcal{X}$ (for transformation d) replace N by N' and C by C'). Here, $\tilde{\mathcal{X}} = \mathcal{D}^{-1/2}(\mathcal{X} - \mathcal{M})$ is the column-wise centered and scaled data matrix where $\mathcal{D} = \operatorname{diag}(\hat{\sigma}_1^2, \ldots, \hat{\sigma}_\kappa^2)$ is the $\kappa \times \kappa$ diagonal matrix of empirical column variances $\hat{\sigma}_l^2$ of $\mathcal{X}$, $l = 1, \ldots, \kappa$, $\mathcal{M} = [\bar{\mathcal{X}}_1, \ldots, \bar{\mathcal{X}}_\kappa] \otimes \mathbf{1}_N$ centers the data matrix by columns with the empirical column means $\bar{\mathcal{X}}_l$, $l = 1, \ldots, \kappa$, and $\mathbf{1}_N$ is a column vector of N ones. The transformation $\tilde{\mathcal{X}}W$ maps the κ reference variables to a new κ-dimensional space. As we use the correlation matrix, the largest variance by scalar projection of $\mathcal{X}$, standardized to variance 1 for each column, lies on the first column of $\tilde{\mathcal{X}}W$, the second largest variance lies on the second column of $\tilde{\mathcal{X}}W$ and so forth up to the smallest variance on the last column (Jolliffe, 2002, p. 30). Finally, we use $W_{1:L}$, the matrix of the first L columns of W, to perform the rank deviation based reference class selection on the dimension reduced matrix obtained as $\mathcal{X}_L = \tilde{\mathcal{X}}W_{1:L}$ for initial transformations a) - c), and $\mathcal{X}_L = \tilde{X}W_{1:L}$ for initial transformation d), where $\tilde{X} = \mathcal{D}^{-1/2}(X - \mathcal{M})$ is the $N \times \kappa$ matrix of the original reference variables column-wise scaled and centered as above. Naturally, we need to calculate $\tilde{\mathcal{X}}_{i,t}W_{1:L}$ for the (potentially) transformed reference variables of the inital firm to assess the rank deviations where $\tilde{\mathcal{X}}_{i,t} = \mathcal{D}^{-1/2}(\mathcal{X}_{i,t} - [\bar{\mathcal{X}}_1, \ldots, \bar{\mathcal{X}}_\kappa]')$ is demeaned as above.

The number of principal components L is chosen by different strategies. On the one hand, for the sake of interpretation we investigate simply using two or three PCs. On the other hand, data driven criteria select the number of PCs that explain at least 75% or 90% of the total variability, or that explain more variability than the mean variability across all PCs Var_μ.

3.3. On Performance Measures of Distributional Forecasts

The resulting distributional information from a reference class serve as forecasts and the suitability of reference classes is assessed by the distributional forecast accuracy. Typically, forecast performance is evaluated by measuring the distance between a forecast and the realized outcome according to a loss function, taking the average loss across all forecast instances and comparing forecast models by their mean loss. Distributional forecasting renders this method infeasible as the realized outcome is not a cumulative distribution function and a distance to the forecast cannot be calculated. We need to evaluate the forecast performance with measures for this specific setting and place the measure Δ_{q} from Chapter 2 in the literature thereon. In line with the prequential principle (Dawid, 1984) we base the evaluation of the forecast model only on forecasts it actually performed and the subsequent realized outcomes in a backtest on historical data.

Here, we must evaluate the forecast quality based on the forecast distribution $F^*_{i,t;h} := n^{-1} \sum_{(j,s)\in R} \mathbb{1}\{Y_{j,s+h} \leq y_{i,t+h}\}$ and the observed outcome $y_{i,t+h}$ of $Y_{i,t+h}$ with distribution function $F_{i,t;h}$ for a fixed forecast horizon h. As we used in Chapter 2, Dawid (1984) and Diebold et al. (1998) propose to use the transformation

$$F^*_{i,t;h}(y_{i,t+h}) = n^{-1} \sum_{(j,s)\in R} \mathbb{1}\{Y_{j,s+h} \leq y_{i,t+h}\} \approx F_{i,t;h}(y_{i,t+h}) \qquad \text{(2.1 revisited)}$$

for forecast evaluation, where $n = |R|$. For an ideal forecast $F^*_{i,t;h} = F_{i,t;h}$, (2.1) holds exactly and $F^*_{i,t;h}(y_{i,t+h})$ is the probability integral transform (PIT) and thus uniformly distributed on $[0, 1]$. Assuming a good forecast, (2.1) should at least hold approximately which makes a near uniform distribution of $F^*_{i,t;h}(y_{i,t+h})$ a necessary condition for a good forecast.

Repeatedly obtaining $F^*_{i,t;h}(y_{i,t+h})$ in a backtest for multiple individuals i and points in time t results in a sample of PIT values $\{p_k\}_{k=1,\ldots,m}$ in $[0, 1]$, where m is the number of forecast instances. If the distribution forecast is valid, (2.1) implies approximate realizations from a uniform distribution on $[0, 1]$. The PIT is useful for absolute assessment whether a predictive distribution is suitable by diagnosing misspecification (Diebold et al., 1998; Gneiting et al., 2007; Held et al., 2010) because uniformity of the PIT values is essentially calibration (probabilistic calibration in Gneiting et al., 2007) and refers to the statistical consistency between observations and the respective distributional forecast. To assess distributional forecast ability, we measure how close this approximation is by checking if it is reasonable to maintain the hypothesis that $\{p_k\}_{k=1,\dots,m}$ stem from a uniform distribution.

The absolute quantile difference Δ_{q} used in Section 2.2.2 compares the $\{\alpha_j\}_{j=1,\dots,l}$ quantiles of $\{p_k\}_{k=1,\dots,m}$ and of the uniform distribution on $[0,1]$, is bounded and enables us to easily calculate an interpretable mean deviation from the theoretical quantiles but is not adjusted for sample size. Admittedly, independent of the number of quantiles there are plenty of distributions that have the same quantiles as the uniform distribution. However, we are more interested in giving suitable reference classes and if a practitioner is particularly interested in certain quantiles of the distribution and not so much in anything else, the absolute quantile difference is feasible. The free choice of quantile levels permits a flexible approach to highlight certain areas of the distribution that researchers or forecasters are interested in most. A visual inspection of histograms of the PIT values is a common forecast assessment (Hamill, 2001) and equivalent to the absolute quantile difference with bins chosen according to the quantile levels. But, while the PIT histogram might be handy if a forecaster only considers a handful of models and/or variable sets, the visual inspection remains qualitative in nature and is infeasible for large scale model and variable selections.

As in Section 2.2.2, statistical goodness-of-fit tests for uniformity are not applicable in this particular backtest as even the smallest forecast errors cause very small p-values due to the large sample sizes. Reporting the value of Kolmogorov-Smirnov (KS) and Cramer-von-Mises (CvM) test statistics avoids this problem and offers a different perspective on the complete distribution in addition to the absolute quantile difference. However, the counterpart to a quantile comparison would be a χ^2 goodness-of-fit test. Tests for continuous distributions might not be suitable if, by construction, the distribution forecast is based on the same number of observations for each forecast instance such that $\{p_k\}_{k=1,\dots,m}$ has a discrete distribution. In such cases and if $Y_{i,t+h}$ is discrete, a χ^2-test is more suitable to assess performance.

Given the difficulty of forecasting corporate sales growth we are mainly interested in finding calibrated distributional forecasts if any exist and do not focus on maximizing the sharpness subject to calibration by sharpness measures or proper scoring rules. Sharpness refers to the concentration of the forecast characterized by scale parameters, is a property of the distributional forecast itself and can be measured by the distances between certain quantiles of the distribution, boxplots, or scale parameters (Bremnes, 2004). A scoring rule is a real-valued function that assigns a loss to a probabilistic forecast $F^*_{i,t;h}$ if the value

Table 3.1.: Hyper parameters for general algorithms in reference class selection.

Name	Abbreviation	Description
reference variables	ref.var.	see Tables 2.1
class size	c	relative size $\in \{0.050, 0.025, 0.010\}$
window length	w	number of past years $\in \{5, 10, 20, 30\}$

$y_{i,t+h}$ is observed (Gneiting et al., 2007). If a scoring rule is proper, the true distribution has a smaller loss than any other forecast distribution such that we check the equality of forecast distributions to the true distribution (Diebold et al., 1998). Scoring rules are suitable for comparative assessment of multiple forecasting schemes if they refer to exactly the same set of forecast situations (Gneiting and Raftery, 2007). Given the vast number of different reference variables and overall frequency of missing values in our data set (see Table 2.2, e.g., more than 50% of the data for 10-year sales growth) providing the same set of forecast instances for each set of reference variables would distort the data set systematically and lead to potentially biased conclusions in the backtest due to survivorship influence. Thus, scoring rules are infeasible here.

3.4. Backtesting Multiple Reference Variables

This backtest evaluates the performance of the new rank-based methods to construct reference classes of sales growth rate forecasts on the firm data set used in Chapter 2 ranging from 1950 to 2019 for forecast horizons one, three, five and 10. Apart from distributional reference class forecasts based on the novel rank deviations and PCA rank deviations, we include the market climate reference class forecast and compare these to results from Chapter 2. Algorithm parameters are shown in Tables 3.1 and 3.2 and, for fixed reference variables, the rank deviation method has 60 possible option combinations and there are 1,200 possible combinations for PCA rank deviation. Backtesting as a special case of cross-validation in time series settings is out-of-sample by construction.

Firms i at time t from the data set are initial firms in the backtest if they are observed at time $t + h$, all used reference variables are observed and $1950 + w + h - 1 \leq t \leq 2019 - h$. Hence, h-year future sales growth and the full timeframe of candidates are available. For a fixed t, all firms j at times s serve as candidates for the initial case's reference class if they are within the window period of candidates, that means, $t - h - w + 1 \leq s \leq t - h$, and if the reference variables and h-year sales growth are available (see Figure 3.3).

Table 3.2.: Different algorithm options and parameters (see Table 3.1). Param. is parameters, #PC is the criterioin to choose principal components and MC is Mauboussin and Callahan (2015). Combination methods union and intersection additionally offer to correct the reference class size by the number of reference variables. Transformation for PCA rank deviation is the pre PCA transformation, the subsequent transformation uses ranks.

Algorithm	Param.	Ref.Var.	Transformation	Combination	#PC
market climate	w	-	-	-	-
group approach	w	SIC	first two or three digits	-	-
MC	w	sales	-	-	-
rank deviations	w, c	all	ranks	LARD, union, intersection	-
PCA rank deviations	w, c	all	none, ranks, trim, $x^{1/5}$	LARD, union, intersection	2, 3, 75% 90%, Var_μ

Thus, for given reference variables and forecast horizon, the data set is restricted to all observations without missing values of these variables where the union of single reference classes in the rank deviation procedure is the exception to the rule (see Section 3.2.2.1). Depending on the set of candidates, the size parameter c and the algorithm we construct a reference class of at least 20 elements. Throughout, we assume that at time t all information of the financial year is available.

For each initial firm (i, t) we obtain a reference class, derive the ECDF of sales growth rates of the reference class elements $\{y_{j,s+h}\}_{(j,s)\in R}$ and the PIT value (2.1). In total, all inital cases produce a sample of PIT values $\{p_k\}_{k=1,\dots,m}$ where the sample size m depends on the forecast horizon h, the window length w, the algorithm and the availability of reference variables. As our main measure of accuracy, we calculate Δ_q based on the PIT values for quantile levels 1%, 5%, 10%, 25%, 50%, 75%, 90%, 95% with $0 \leq \Delta_\text{q} \leq 4.5$, here. The choice of quantiles is motivated by an emphasis on the distribution tails in contrast to a set of equidistant quantile levels. Further, we report KS and CvM test statistics as accuracy measures of the whole distributional approximation.

3.4.1. Variable and Model Selection Procedure

Finding appropriate reference classes is in essence a variable and model selection problem. We systematically explore which reference variables contain information for a calibrated distributional forecast based on rank deviations. To this end, we use a forward selection

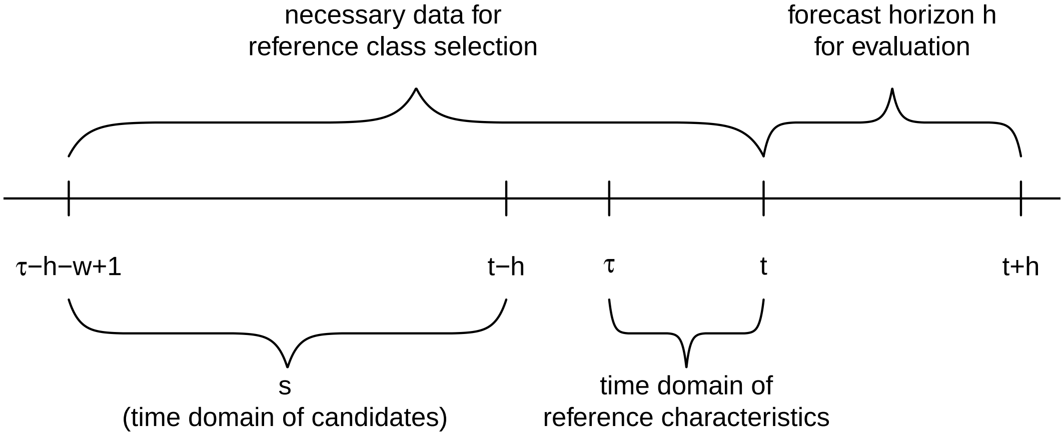

Figure 3.3.: Illustration of the backtest timeline. Note, $\tau \leq t - h$ is possible as well.

and brute force approaches on all contemporaneous reference variables and on selected reference variable subsets. PCA rank deviation applied on selected reference variable subsets completes the procedure and we backtest $67,080$ different variable and rank deviation options for each forecast horizon.

For a systematic backtest of the rank deviation algorithm on multiple reference variables we use a forward selection to reduce the number of possible reference variable combinations. We begin with the best three reference variables according to Δ_q from results in Chapter 2 for each forecast horizon and combine them with each of the remaining reference variables using 60 different algorithm options. We continue with the three best reference variable pairs from the previous stage with two reference variables and combine them with each of the remaining reference variables and all possible algorithm options. Then, we repeat this for every stage by choosing the three best sets of reference variables from the previous stage. We stop if adding another reference variable does not improve the results anymore. However, if the forward selection comes to an early halt we continue in order to protect against finding a local minimum. The forward selection terminates when results for none of the forecast horizons improve, that is, after using six reference variables. Thus, we backtest $21,780$ different combinations.

Further, we explore using exclusively contemporaneous reference variables due to the lower data requirements opposed to lagged variables that are chosen by the forward selection. Therefore, we brute force all combinations of seven contemporaneous reference variables (except SIC) and additionally combine the balance sheet variables (i.e., *sales*, *operating margin*, *total assets* and *shareholder equity*) and the full set of contemporaneous variables with up to one-, three-, five- and 10-year *past sales growth* and *past operating margin delta*, respectively. This results in 127 variable combinations in the brute force approach and eight sets of variables with different degrees of lagged variables and, thus, $8,100$ different combinations for each forecast horizon.

Finally, we investigate the influence of dimension reduction on forecast performance using PCA on different sets of reference variables before applying the rank deviation methods. The 31 reference variable subsets under consideration are all four balance sheet variables, all contemporaneous variables, each of these variable sets combined with up to one-, three-, five- and 10-year lagged variables, the combination of the four, five and six best reference variables, respectively, from the single reference variable approach, and each at a time the three best sets of four, five, and six best reference variables from the brute force approach and from the forward selection.[6] This results in $37,200$ different combinations for each forecast horizon.

3.4.2. Results of Backtesting Multiple Reference Variables

Tables 3.3 - 3.6 show a selection of our results[7] on forecast horizons one, three, five and 10 years, each ranked by Δ_q to compare the novel methods to results from Chapter 2. The rank deviation (RD) results reported are the three best overall combinations, the best combination of contemporaneous reference variables and for both, in view of the necessary data, the best combination using a five- and 10-year window, respectively. For PCA rank deviation (PCARD) we show the same selection of results as for RD. We additionally report the best market climate window, the best results for the group approach, the best window for the method in Mauboussin and Callahan (2015) and the best single reference variable overall as well as for a five- and 10-year window. Some of these cases coincide, thus, each table has at most 22 rows. We give details on algorithm options according to Tables 3.1 and 3.2 and on reference variables (see Table 2.1).[8]

Across all forecast horizons the algorithms using several reference variables improve distributional forecast performance and reduce Δ_q by between 38% and 71%. This implies that the mean error for quantile levels of the distribution of, for example, one-year

[6] Results for these variable subsets can be found in Tables 3.9 - 3.16 in Appendix 3.A.

[7] Full results are available upon request.

[8] Here, contemp. refers to all contemporaneous reference variables and in subscript $s{:}t$ is $\{s, \ldots, t\}$.

Table 3.3.: Comparison of reference variables (ref. var.) and algorithms for forecasting one-year ahead sales growth. Alg. is algorithm, transf. is pre PCA transformation, comb. is combination of reference variables and cor. is correction.

Alg.	Ref. Var.	Transf.	#PC	Comb.	Cor.	w	Size	Δ_q	KS	CvM
PCA	contemp., salesGR$_1$, opmarΔ_1	ranks	3/ Var_μ	union	yes	30	0.01	0.0045	1.2002	0.1703
PCA	contemp., salesGR$_1$, opmarΔ_1	ranks	3/ Var_μ	union	yes	30	0.01	0.0046	1.3626	0.1857
PCA	contemp., salesGR$_1$, opmarΔ_1	ranks	3/ Var_μ	union	yes	30	0.025	0.0047	1.4200	0.1755
RD	salesGR$_{3,5:7}$, opmarΔ_5	–	–	union	no	30	0.05	0.0065	0.7191	0.0407
RD	salesGR$_{5:7}$, opmarΔ_5	–	–	union	no	30	0.01	0.0066	0.7408	0.0448
RD	salesGR$_{5:8}$, opmarΔ_5	–	–	union	no	30	0.05	0.0072	0.6475	0.0698
PCA	contemp., salesGR$_1$, opmarΔ_1	ranks	2	union	no	10	0.01	0.0094	1.6183	0.5917
PCA	contemp.	trim	75%	union	no	30	0.025	0.0102	1.2817	0.1950
single	opmarΔ_6	–	–	–	–	30	0.025	0.0157	1.8644	0.8265
RD	β, P/E	–	–	union	no	5	0.025	0.0158	2.7215	1.8060
PCA	contemp., salesGR$_1$, opmarΔ_1	ranks	2	unio	yes	5	0.025	0.0164	2.3868	1.3622
RD	β, P/E	–	–	union	no	10	0.05	0.0213	3.9850	4.9765
PCA	sales, at, seq, P/E, P/B	$x^{1/5}$	90%	union	no	10	0.025	0.0217	2.6242	2.4904
PCA	sales, opmar, at, seq	trim	2/ 75%	union	yes	5	0.025	0.0233	5.6027	7.1741
single	opmar	–	–	–	–	10	0.05	0.0284	4.1500	6.1454
single	opmar	–	–	–	–	5	0.05	0.0309	4.4533	4.8490
market	–	–	–	–	–	5	–	0.0454	6.0073	11.1804
MC	sales	–	–	–	–	5	–	0.0516	6.3825	12.7518
group	major group	–	–	–	–	5	–	0.0653	8.6576	22.5482

sales growth reduces on average to 0.05 percentage points when predicting on historical data. Generally, the reduction is greater with shorter forecast horizons and overall the results improve with a shorter forecast horizon. For all forecast horizons the best results are delivered by PCARDs with a fixed number of PCs based on contemporaneous reference variables combined with different degrees of lagged *past operating margin deltas* and *past sales growth rates.* However, past reference variables do not exceed a lag of five years. All reference variables should be either 2.5% trimmed on both tails before PCA or one should use ranks for PCA. While combining reference variables by union, a window length of 30 years is best overall, besides a 20-year window for three-year forecast horizon. Notably, the choices of reference class size and correction vary across the best results. Under some constraints, LARD and intersection are among the best combination versions and the number of PCs gets chosen by a data driven criterion although the overall best results still use a fixed number of PCs and union for combination.

All contemporaneous variables are used for the best combination and under certain constraints the market parameters β, *price-to-earnings ratio* and *price-to-book ratio* are important. An exception is the 10-year forecast horizon where only balance sheet variables are selected with a possible interpretation that market parameters better reflect short term expectations. The importance of *past operating margin deltas* supports the results in Section 2.4.1 and is discussed there. However, the excellent performance of *past sales growth rates* contradicts the part of Gibrat's law that claims growth rates are uncorrelated in time (Gibrat, 1931). Stanley et al. (1996) also show that sales growth depends on past growth rates and that sales growth distributions are similar across diverse firms which corresponds to the poor results of the group approach here.

Using multiple reference variables with RD improves the results by between 58% and 12% compared to the single variable use. The *past operating margin deltas* and *past sales growth rates* dominate the forward selection with lags mainly between three and eight years, partially up to 10 years, a window length of 30 years and the number of used reference variables in the best combination varies from two to five (c.f. Tables 3.9, 3.11, 3.13 and 3.15 in Appendix 3.A). Indeed, taking more reference variables into consideration does not yield better results in general. This seems to be a feature of the specific rank-based algorithms used here as, for example, combining contemporaneous variables with *past sales growth rates* and *past operating margin deltas* up to different lags performs substantially worse compared to single reference variables where Δ_{q} is between 1.7 and 3.7 times higher depending on forecast horizon (see Table 3.8 in Appendix 3.A). However, this may partially explain the outstanding performance of PCARD using a small fixed number of PCs. As for PCARD, the best RD options vary across size and correction but all combine the reference variables by union. To put this into perspective, union means that the reference class members are similar to the initial firm in at least one reference variable in contrast to similarity across variables when using LARD or intersection. This superiority reflects the fact that the outside view, often seen as relying on *only superficially similar instances*, produces more accurate predictions than a narrow-minded focus on the uniqueness and complexity of a forecast challenge (c.f. Lovallo and Kahneman, 2003). The benchmarks of market climate and group approach as well as of Mauboussin and Callahan (2015) are clearly outperformed across all forecast horizons. There are some different rankings of algorithms across accuracy measures which is natural as there is no universal ranking of forecasts regardless of the accuracy measure (see Diebold et al., 1998) and given the focus of KS and CvM on the complete distribution.

With practical application in mind the amount of necessary data is important and consists of two components, namely the years used to select a window of candidates and the lags of *past sales growth* and *past operating margin delta*. Collecting a smaller data basis is easier to achieve in practice and further takes into account that practitioners might want to assume a stable data generating mechanism of sales growth for only a few years. However, smaller windows generate smaller candidate sets and ultimately smaller reference classes in general. In particular, *past operating margin deltas* and *past sales growth rates* turn out to be best in the forward selection but depend on a rather large amount of data. The best performances by PCARDs need between 23 and 35 years of data while the best RD and the best single variable combinations need between 36 and 40 years of data despite performing worse. Remarkably, the use of dimension reduction via PCA enables us to incorporate more reference variables while simultaneously needing less data and improving the results substantially.

Further, the following results in Tables 3.3 - 3.6 stand out, where changes are reported with respect to the best single reference variable for the respective forecast horizon. For one-year forecast horizon (see Table 3.3) PCARD improves Δ_q by 71% and RD improves Δ_q by 58% while using five years less and one year more of data, respectively. There is even a 39% improvement for PCARD with 25 years of necessary data less. Notably, RD using β and *price-to-earnings ratio* performs roughly the same as the best single reference variable but needs only five instead of 36 years of data. Interestingly, for the three best PCARD combinations the results for three PCs and all PCs that explain at least the mean variance are identical. When forecasting three years ahead, Table 3.4

Table 3.4.: Comparison of reference variables (ref. var.) and algorithms for forecasting three-year ahead sales growth. Alg. is algorithm, transf. is pre PCA transformation, comb. is combination of reference variables, cor. is correction and inters. is intersection.

Alg.	Ref. Var.	Transf.	#PC	Comb.	Cor.	w	Size	Δ_q	KS	CvM
PCA	contemp., salesGR$_{1:3}$, opmar$\Delta_{1:3}$	trim	3	union	yes	20	0.01	0.0146	2.2379	1.3913
PCA	contemp., salesGR$_{1:5}$, opmar$\Delta_{1:5}$	trim	3	union	yes	20	0.01	0.0163	1.9873	1.1432
PCA	contemp., salesGR$_{1:5}$, opmar$\Delta_{1:5}$	trim	3	union	no	20	0.05	0.0165	2.2956	1.4306
RD	salesGR$_{10}$, opmar$\Delta_{8,9}$	–	–	union	no	30	0.01	0.0223	1.7135	0.7237
RD	salesGR$_{10}$, opmar$\Delta_{8,10}$	–	–	union	yes	30	0.01	0.0233	1.7000	0.6886
PCA	sales, opmar, seq, β, P/E, P/B	trim	Var$_\mu$	union	yes	30	0.01	0.0238	3.1099	2.3311
RD	salesGR$_{10}$, opmarΔ_8	–	–	union	yes	30	0.01	0.0239	1.4898	0.5776
PCA	opmar$\Delta_{6:8,10}$	–	75%	inters.	no	10	0.01	0.0282	1.3639	0.3202
single	opmarΔ_7	–	–	–	–	30	0.025	0.0290	3.2390	2.8895
RD	β, P/E, P/B	–	–	LARD	–	10	0.05	0.0319	4.7188	6.6191
RD	β, P/E, P/B	–	–	LARD	–	5	0.025	0.0334	4.0785	3.7155
PCA	opmar$\Delta_{6:10}$	–	75%	inters.	no	5	0.025	0.0373	3.0515	2.6985
PCA	sales, opmar, β, P/E	$x^{1/5}$	2	union	yes	5	0.05	0.0460	6.6272	12.6316
PCA	sales, opmar, seq, β, P/E, P/B	$x^{1/5}$	2	union	yes	10	0.05	0.0486	6.5503	12.6209
single	opmar	–	–	–	–	30	0.01	0.0603	6.9167	16.7652
single	opmar	–	–	–	–	5	0.05	0.0707	10.4687	33.6970
single	opmar	–	–	–	–	10	0.05	0.0883	11.8219	55.6199
market	–	–	–	–	–	5	–	0.0924	11.3359	45.3895
MC	sales	–	–	–	–	5	–	0.1028	13.4856	61.3178
group	major group	–	–	–	–	5	–	0.1423	17.9423	106.9768

Table 3.5.: Comparison of reference variables (ref. var.) and algorithms for forecasting five-year ahead sales growth. Alg. is algorithm, transf. is pre PCA transformation, comb. is combination of reference variables, cor. is correction and inters. is intersection.

Alg.	Ref. Var.	Transf.	#PC	Comb.	Cor.	w	Size	Δ_q	KS	CvM
PCA	contemp., salesGR$_{1:5}$, opmar$\Delta_{1:5}$	trim	2	union	yes	30	0.01	0.0179	1.0948	0.2297
PCA	contemp., salesGR$_{1:5}$, opmar$\Delta_{1:5}$	trim	2	LARD	–	30	0.01	0.0186	1.5236	0.8604
PCA	contemp., salesGR$_{1:5}$, opmar$\Delta_{1:5}$	trim	2	union	yes	30	0.025	0.0207	1.1641	0.3481
RD	salesGR$_{10}$, opmarΔ_6	–	–	union	yes	30	0.01	0.0230	2.2279	1.0487
RD	salesGR$_{10}$, opmarΔ_6	–	–	union	no	30	0.01	0.0261	2.1110	1.0450
PCA	sales, opmar, seq, β, P/E, P/B	trim	90%	union	no	30	0.05	0.0261	3.0871	1.7590
RD	salesGR$_{6,10}$, opmarΔ_6	–	–	union	yes	30	0.01	0.0264	1.5280	0.6581
single	opmarΔ_{10}	–	–	–	–	30	0.01	0.0320	2.2045	1.3087
PCA	sales, opmar, at, seq, salesGR$_{1:5}$ opmar$\Delta_{1:5}$	trim	Var$_\mu$	inters.	no	5	0.025	0.0394	1.3993	0.1071
PCA	opmar$\Delta_{4:7,9,10}$	–	Var$_\mu$	inters.	no	10	0.025	0.0399	4.2461	5.1859
PCA	sales, opmar, β, P/E, P/B	trim	75%	inters.	yes	5	0.025	0.0484	1.2470	0.1948
RD	β, P/E	–	–	LARD	–	30	0.05	0.0492	5.1695	6.0082
PCA	sales, opmar, β, P/E, P/B	trim	2	inters.	no	10	0.01	0.0559	1.7647	0.3353
RD	opmar, opmarΔ_6	–	–	inters.	yes	5	0.01	0.0634	2.1629	0.4439
RD	opmar, opmarΔ_{10}	–	–	inters.	yes	10	0.01	0.0794	2.4536	0.3912
RD	β, P/E, P/B	–	–	LARD	–	5	0.05	0.0716	6.7605	12.5422
single	opmar	–	–	–	–	30	0.01	0.0856	9.4768	32.0558
RD	opmar, P/B	–	–	inters.	no	10	0.05	0.0863	1.3772	0.3369
single	P/E	–	–	–	–	5	0.05	0.1113	9.1733	41.2639
market	–	–	–	–	–	5	–	0.1428	15.2365	96.8583
single	P/E	–	–	–	–	10	0.05	0.1497	13.1570	84.6791
MC	sales	–	–	–	–	5	–	0.1600	19.0380	137.3940
group	major group	–	–	–	–	30	–	0.2136	16.7058	106.9918

Table 3.6.: Comparison of reference variables (ref. var.) and algorithms for forecasting 10-year ahead sales growth. Alg. is algorithm, transf. is pre PCA transformation, comb. is combination of reference variables, cor. is correction and inters. is intersection.

Alg.	Ref. Var.	Transf.	#PC	Comb.	Cor.	w	Size	Δ_{q}	KS	CvM
PCA	sales, opmar, at, seq, salesGR$_{1:5}$ opmar$\Delta_{1:5}$	trim	2	union	yes	30	0.01	0.0275	1.5735	0.6993
PCA	sales, opmar, at, seq, salesGR$_{1:5}$ opmar$\Delta_{1:5}$	trim	2	LARD	–	30	0.01	0.0284	1.7164	0.8456
PCA	sales, opmar, at, seq, salesGR$_{1:5}$ opmar$\Delta_{1:5}$	trim	2	union	yes	30	0.025	0.0296	1.9586	0.7830
PCA	sales, opmar, at, seq, β	trim	Var$_{\mu}$	inters.	no	20	0.01	0.0323	1.5376	0.4871
RD	salesGR$_{6,7}$, opmar$\Delta_{5,7}$	–	–	union	yes	30	0.01	0.0388	3.0820	3.2019
RD	salesGR$_{6}$, opmarΔ_{7}	–	–	union	yes	30	0.01	0.0401	3.1193	3.2598
RD	salesGR$_{6:8}$, opmar$\Delta_{5,7,8}$	–	–	union	yes	30	0.01	0.0403	2.9479	2.9308
single	opmarΔ_{6}	–	–	–	–	30	0.025	0.0441	3.7773	4.1454
PCA	opmar, β, P/E, P/B	trim	90%	inters.	no	5	0.05	0.0584	1.1665	0.2001
RD	β, P/E	–	–	LARD	–	30	0.025	0.0679	4.1868	4.2020
PCA	opmar, β, P/E, P/B	trim	3	inters.	yes	10	0.01	0.0705	1.2347	0.4220
RD	β, salesGR$_{5}$, opmarΔ_{7}	–	–	LARD	–	10	0.05	0.0785	6.0266	10.9679
RD	salesGR$_{7}$, opmar$\Delta_{3,5,6}$	–	–	inters.	no	5	0.025	0.0998	1.0023	0.1382
single	opmar	–	–	–	–	30	0.01	0.1126	7.4249	20.6290
RD	opmar, β, P/E	–	–	LARD	–	10	0.05	0.1147	8.6451	17.1154
RD	opmar, β, P/E	–	–	LARD	–	5	0.05	0.1154	9.5327	24.3167
single	opmarΔ_{10}	–	–	–	–	10	0.025	0.2053	8.5536	31.8502
MC	sales	–	–	–	–	30	–	0.2270	11.2416	50.6546
group	major group	–	–	–	–	30	–	0.2561	12.0198	61.4773
market	–	–	–	–	–	30	–	0.2845	14.4029	74.5287
single	opmarΔ_{10}	–	–	–	–	5	0.025	0.2926	14.5939	94.3775

shows a reduction of Δ_{q} by 50% and by 23% for PCARD and RD with 14 years less and three years more of data, respectively. Even PCARD using 17 years of information less is slightly better and additionally, PCARD on contemporaneous variables only with a 30-year window is better than the single variable method as well. Moreover, findings on five-year ahead forecasting in Table 3.5 show an improvement of Δ_{q} by 44% and by 28% for PCARD and RD while needing five years less and the same amount of past data. In contrast, a PCARD combination with contemporaneous variables and a 30-year window improves results by 18% while needing 10 years of data less. Finally, in Table 3.6 on 10-year forecast horizon, contemporaneous balance sheet variables and past variables are best for PCARD and reduce Δ_{q} by 38% needing similar years of data. While the best RD combinations have a comparable data demand as the single variable approach and improve Δ_{q} by 12%, notably, a PCA combination using intersection on only contemporaneous variables from a 20-year window improves results by 27% despite using 16 years of information less.

3.5. Practical Application

In any application, a distributional forecast can be used to provide a comparison to existing forecasts and prediction intervals, point forecasts or probabilities for intervals of possible outcomes can be directly calculated from the ECDF of the reference class. Point forecasts, for example, model based or by experts, are assessed by calculating the PIT value $\mathbb{P}(Y_{i,t+h} \leq y_{i,t+h}) \approx n^{-1} \sum_{(j,s)\in R} \mathbb{1}\{Y_{j,s+h} \leq y_{i,t+h}\}$ for $n = |R|$. PIT values close to either 0 or 1 can serve as a warning signal to check for arguments that may justify the forecast relative to the reference class or possibly correct the prediction. Here, we provide an application of the reference class approach on forecasting sales growth over multiple years and additionally assess expert forecasts. The results are compared to those from Chapter 2.

Here, the distributional reference class forecasts of sales growth based on multiple variables for the example firms 3M and Amazon show similar results compared to the distributional forecasts based on a single variable from Section 2.5. For both companies, the distributional one-year sales forecasts based on year 2018 are compared to analysts' forecast from the FactSet (2021) estimates database and Figures 3.4 and 3.5 present density functions in comparison to these estimates. The reference classes for one-year sales growth forecasts in this section are constructed according to the best result of our backtest in Table 3.3 using all contemporaneous variables, *one-year past sales growth* and

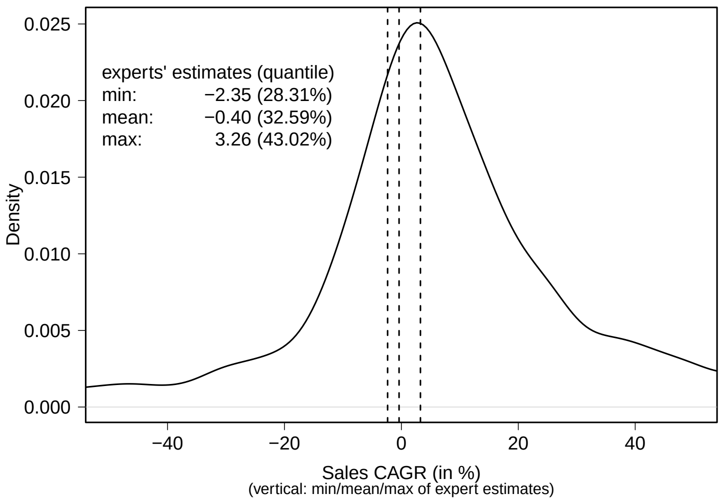

Figure 3.4.: Forecasted density of one-year sales growth for 3M based on the best algorithm options from Table 3.3 compared to experts' estimates. Density estimation on support $[-100, \infty)$ is based on the Gaussian kernel and Silverman's rule of thumb provides the bandwidth.

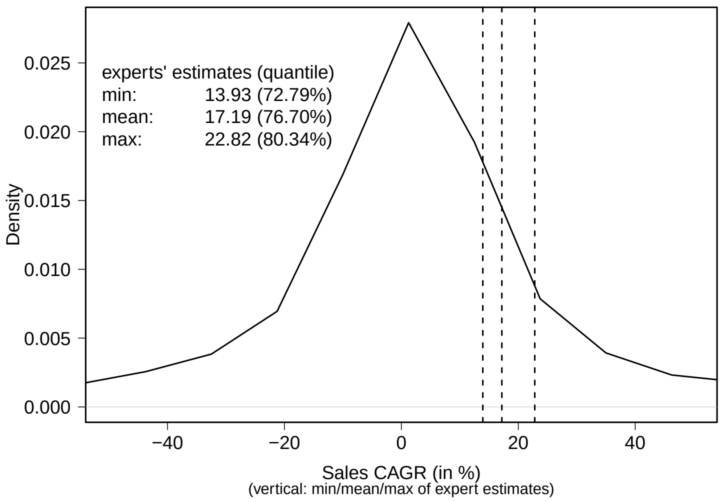

Figure 3.5.: Forecasted density of one-year sales growth for Amazon based the best algorithm options from Table 3.3 compared to experts' estimates. Density estimation on support $[-100, \infty)$ is based on the Gaussian kernel and Silverman's rule of thumb provides the bandwidth.

one-year past operating margin delta as reference variables. The selection is based on a PCA rotation of the reference variables' ranks with 3 PCs where candidates are chosen from a 30-year window period. All reference classes for individual PCs have size 0.33% of the candidate set due to the size correction and, then, the three reference classes are unified.

For 3M, expert forecasts range from the 28.31% to the 43.02% quantile within the reference class and, similar to the single variable approach, indicate no sign of overoptimism. Inside and outside view roughly agree for the new method as well and classify 3M as an average company regarding sales growth. However, the reference class predicts a probability of only 14.71% that sales growth lies within the range of expert forecasts. Thus, the outside view can protect against overconfidence in expert forecasts. In case of Amazon, forecasts correspond to quantiles of the distributional forecast between 72.79% and 80.34% and are more optimistic than for 3M since only one out of five firms in the reference class achieved the maximum predicted growth of Amazon. The new method indicates that the expert forecasts are not as optimistic as the single variable method suggests. However, such a result should still prompt forecasters to justify or correct their predictions, even though Amazon is known for its capability of high growth. This way the outside view may yield protection against extreme and unrealistic predictions

To put the size of the reference classes into context, there are 271,548 firm observations before 2018 available in the data set and restricting them to all firms providing the necessary reference variables in a 30-year window shrinks these to a set of 109,792 candidate firms. Out of these candidates, the algorithm chooses roughly 0.98% of the candidates or 1,074 and 1,073 observations for the reference classes of 3M and Amazon, respectively. To put this into perspective, the best combination with a 10-year window in Table 3.3, a set of options that needs 20 years of data less, operates on a set of 27,346 candidates and selects only about 1.98% of the candidates or 541 and 542 observations, respectively. Note that the former algorithm option used a size correction and the latter not, resulting in a reference class twice as big as the size parameter suggested as there are few observations that are chosen according to both PCs.

Table 3.7 shows base rates for one-, three-, five- and 10-year forecasting horizons for 3M and Amazon and underlines the higher growth chances of Amazon. We demonstrate the usefulness of these base rates by considering the example of an entity that wants to invest their money in rising businesses that have the highest probability of a long term sales growth above 5% per year. Assuming 10 years to be long term, we can directly use

Table 3.7.: Comparison of reference classes for forecasting compound annual sales growth rates of 3M and Amazon with base year 2018. The choice of algorithm for each forecast horizon is based on the results of the backtests in Tables 3.3 – 3.6. Mean and standard deviation are 2.5% trimmed on both tails.

	Base Rates							
	1-Yr		3-Yr		5-Yr		10-Yr	
CAGR (%)	3M	Amazon	3M	Amazon	3M	Amazon	3M	Amazon
≤ -25	7.54	10.81	2.70	5.45	1.51	2.57	0.60	1.40
$]-25,-20]$	2.05	2.52	1.98	1.64	1.06	0.76	0.20	0.40
$]-20,-15]$	1.58	3.45	1.98	2.73	2.57	1.81	1.20	1.00
$]-15,-10]$	4.00	3.45	5.41	2.73	3.78	4.23	2.80	3.00
$]-10,-5]$	7.45	7.83	9.19	7.27	9.23	5.74	9.00	7.00
$]-5,0]$	11.17	13.05	17.84	11.45	19.97	14.50	23.80	18.80
$]0,5]$	14.25	16.50	20.36	18.55	28.74	21.75	38.20	27.00
$]5,10]$	10.89	9.23	13.15	13.09	16.34	15.26	14.00	14.80
$]10,15]$	8.94	7.08	9.73	10.18	8.17	12.69	6.00	11.80
$]15,20]$	6.15	4.75	6.67	6.91	4.39	6.50	2.40	5.40
$]20,25]$	4.56	2.61	3.24	4.18	2.12	4.68	0.60	5.20
$]25,30]$	3.91	2.52	0.90	3.27	1.06	3.47	0.60	1.80
$]30,35]$	1.86	1.49	1.44	3.27	0.45	2.72	0.40	1.00
$]35,40]$	2.33	1.21	1.44	2.00	0.30	0.60	0.20	0.40
$]40,45]$	1.86	1.03	1.26	2.00	0.15	0.45	0.00	0.40
> 45	11.45	12.49	2.70	5.27	0.15	2.27	0.00	0.60
mean	10.63	11.98	4.13	7.16	2.23	6.16	1.63	4.57
median	5.99	2.73	2.81	5.08	2.36	4.65	1.38	3.21
std	26.40	47.85	12.57	16.95	8.09	11.55	5.85	8.75

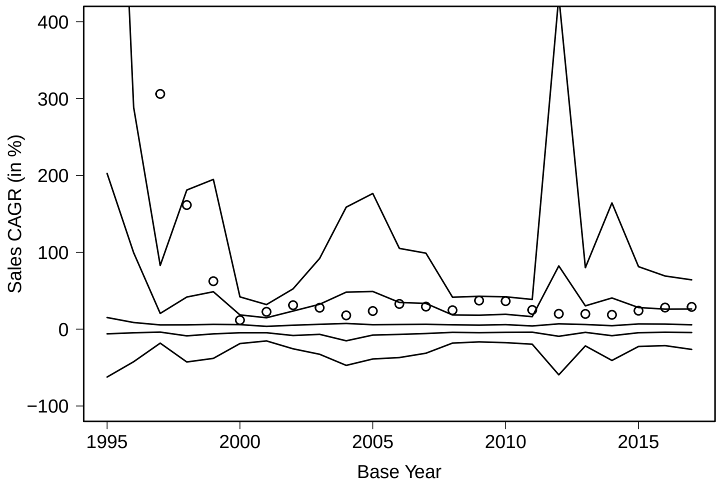

Figure 3.6.: One-year sales growth for Amazon from 1995 to 2017 compared to quantiles of the reference class outcomes selected using the best algorithm options from Table 3.3. The bold lines from bottom to top represent the 10%, 25%, 50%, 75% and 90% quantiles of one year sales growth within the reference class. The circles represent sales growth of Amazon which is 2,891% and 823% for base years 1995 and 1996, respectively, and therefore not displayed in this graph.

the base rates in Table 3.7 to predict such a probability for both companies by adding up the relevant cells in the most right columns. That results in predicted probabilities of 24.2% and 41.4% for more than 5% compound annual sales growth for 3M and Amazon, respectively. For all forecast horizons, especially the base rates for larger sales growth are higher for Amazon than for 3M, and, in general, the predicted distribution has a higher variability for Amazon. This may be interpreted as the higher risk that contemplates the higher potential reward. Overall, the standard deviation declines with the forecast horizon as we display compound annual growth rates. Interestingly, for one-year forecast horizon the base rates for 3M show less probability for sales decline than for Amazon and, except for growth above 45%, 3M has higher base rates for sales increase. This contemplates Figure 3.4 where the expert forecasts are roughly centered at unchanged sales but the distributional reference class forecast shows a tendency to sales increase. However, the base rates for one-year sales growth exhibit by far the most uncertainty, as can be seen by the trimmed standard deviation, and also the highest predicted probabilites for sales growth exceeding 45% and more than 25% probability of sales decline.

Moreover, we stick to the example of Amazon and present another useful application of reference classes. Figure 3.6 follows Amazon's one year ahead distributional sales growth forecasts through the base years 1995 to 2017 and compares these forecasts to the realized sales growth. Here, it shows that Amazon is outperforming its reference class massively in the first three years. After that, the realized sales growth rate is close to the 75% quantile for most of the years, with some fluctuations between the 50% and 90% quantile of its reference classes. With respect to the situation in Figure 3.5 this may serve as a validation of expert forecasts. More general, the dependence of reference class selection on historic data gets evident as the distributional forecasts' uncertainty increases in the aftermath of well-known times of financial distress, here, the dotcom crisis in 2000, the subprime bubble in 2007 and 2008, and the European debt crisis in 2009 and 2010. Overall, a practitioner might conclude Amazon performs well compared to peers and is in a good overall market position. On the one hand, this seems like old news, but on the other hand it serves as an affirmation and a proof of concept: Distributional reference class forecasting is well behaved and meets practical expectations in this case.

3.6. Concluding Remarks

In this chapter, we extend the analysis of distributional reference class forecasting of corporate sales growth with a focus on reference class selection. We provide a practical solution to the well-known reference class problem (Venn, 1888) that arises in any application of reference class forecasting as described in Kahneman and Tversky (1979). The novel rank-based methods enable the use of several reference variables and include the option of a dimension reduction based on principal components. In an extensive backtest on corporate data from the USA covering several decades we conclude that especially principal component analysis reduces the amount of necessary past data while simultaneously improving the results from Chapter 2 substantially by between 38% and 71% depending on forecast horizon. Further, we illustrate the practical usefulness of the new methods by forecasting distributional sales growth for two example firms, 3M and Amazon. The novel approaches need less historic observations compared to existing methods, are easy to interpret and deliver reasonable results making them useful for practical applications. However, there are further extensions possible.

The method itself can be extended by using additional algorithms for reference class selection. Other methods for dimension reductions are possible, for example, the self-organizing map (Kohonen, 1982), an artificial neural network using a two dimensional grid of neurons for dimension reduction. On data sets with a cluster structure we could use the neurons of the self-organizing map as cluster centers or other cluster algorithms. A parametric model for the distribution of sales growth within the reference classes (as in Stanley et al., 1996) could be used for distributional reference class forecasting.

Within our method, it is crucial to rank the forecast ability of the different algorithms and reference variables. We have not answered whether the results on calibration differ statistically significantly between the forecasts. The only indication on acceptable numerical regions of the accuracy measures for generating appropriate reference classes is given by the results of the market climate approach that can be interpreted as a prediction of the marginal distribution. But it would be quite useful to know which forecasts are in fact calibrated. Given a data set with less missing values, we could then use scoring rules that additionally assess the sharpness of forecasts and are more suitable for comparative assessment of distributional forecasts (Gneiting and Raftery, 2007).

It is yet open whether reference classes can identify underlying distributions which could be answered in a simulation study to deepen the unterstanding of the mechanism behind reference class selection. A study on similarity based forecasting using a weighted mean of reference class outcomes to issue point forecasts is also possible (Gilboa et al., 2006). On a similar line of thought, correcting potentially biased expert (or model based) forecasts with outside views is yet to be investigated as the original corrective procedure in Kahneman and Tversky (1979) suggested. This means that expert forecasts are combined with a reference class forecast and a backtest checks for forecast improvement.

Investigations on other data sets beyond the presented case are necessary to further advocate the utility of the approach. Additional possible applications are characterized by availability of sufficient data on past outcomes and by the fact that forecasts should typically be hard to issue. Ideally, no models producing calibrated (distributional) forecasts directly should be known in the literature or existing models should be very complicated and/or not accepted by a broad audience of practitioners and thus sparsely used. In the field of finance, forecasting of cash flow items (in corporate value theory), bankruptcy probabilities (or credit rating) and financial returns to assess value at risk may be possible further applications.

3.A. Supporting Tables

This section contains tables with supporting results of the backtest in Section 3.4. Table 3.8 shows that more reference variables do not necessarily improve reference class selection based on rank deviation (c.f. the discussion in Section 3.4.2). Tables 3.9, 3.11, 3.13 and 3.15 contain the best results of the forward selection of the rank deviation procedure for all considered forecast horizons as described in Section 3.4.1. The best three sets of four, five and six reference variables of the forward selection are used in backtesting PCA rank deviation, see Section 3.4.1. Tables 3.10, 3.12, 3.14 and 3.16 contain the best results of the brute force backtest of rank deviation on contemporaneous reference variables for all considered forecast horizons as described in Section 3.4.1. The best three sets of four, five and six reference variables from the brute force approach are used in the backtest of PCA rank deviation as well, see Section 3.4.1.

Table 3.8.: Best results for combining contemporaneous reference variables with past sales growth rates and past operating margin deltas up to different lags for forecast horizons one, three, five and ten.

Horizon	Ref.Var.	Comb.	Cor.	w	Size	Δ_q	KS	CvM
1	contemp., salesGR$_1$, opmarΔ_1	union	no	30	0.05	0.0584	7.0061	15.9007
3	contemp., salesGR$_1$, opmarΔ_1	union	no	30	0.01	0.0773	8.0135	24.0522
5	contemp., salesGR$_1$, opmarΔ_1	union	yes	30	0.01	0.0886	8.4625	26.1944
10	contemp., salesGR$_{1:5}$, opmar$\Delta_{1:5}$	union	yes	30	0.01	0.0758	4.7809	6.7344

Table 3.9.: Forward selection results for forecasting one-year sales growth. The three best sets of four, five and six reference variable, respectively, are used in the backtest of PCA rank deviation.

Choice	Ref.Var.	Comb.	Cor.	w	Size	Δ_q	KS	CvM
Best single	opmarΔ_6	–	–	30	0.025	0.0157	1.8644	0.8256
	opmarΔ_7	–	–	30	0.025	0.0159	2.2179	1.0808
	opmarΔ_5	–	–	30	0.01	0.0171	2.3873	1.2154
	opmarΔ_9	–	–	30	0.025	0.0187	2.4281	1.1636
	opmarΔ_{10}	–	–	30	0.01	0.0188	2.0381	0.7830
	opmarΔ_3	–	–	30	0.01	0.0202	2.5357	1.6515
Best 2	salesGR$_7$, opmarΔ_5	union	no	30	0.05	0.0114	1.6476	0.5886
	salesGR$_8$, opmarΔ_6	union	no	30	0.05	0.0114	1.5929	0.5830
	salesGR$_5$, opmarΔ_7	union	yes	30	0.05	0.0115	0.9091	0.1290
Best 3	salesGR$_{5,6}$, opmarΔ_7	union	yes	30	0.01	0.0087	0.8496	0.1021
	salesGR$_{5,7}$, opmarΔ_5	union	no	30	0.025	0.0098	0.8259	0.0843
	salesGR$_{5,7}$, opmarΔ_7	union	yes	30	0.01	0.0098	0.9199	0.1943
Best 4	salesGR$_{5:7}$, opmarΔ_5	union	yes	30	0.01	0.0066	0.7408	0.0448
	salesGR$_{5,7,8}$, opmarΔ_5	union	no	30	0	0.0072	0.6538	0.0717
	salesGR$_{3,5,6}$, opmarΔ_7	union	no	30	0	0.0086	0.8131	0.1163
Best 5	salesGR$_{3,5:7}$, opmarΔ_5	union	no	30	0.05	0.0065	0.7191	0.0407
	salesGR$_{5:8}$, opmarΔ_5	union	no	30	0.05	0.0072	0.6475	0.0698
	salesGR$_{3:6}$, opmarΔ_7	union	no	30	0.05	0.0084	0.6955	0.0546
Best 6	salesGR$_{3:7}$, opmarΔ_5	union	no	30	0.05	0.0076	0.7212	0.0628
	salesGR$_{3:7}$, opmarΔ_7	union	no	30	0.025	0.0093	0.7668	0.1116
	salesGR$_{5:8}$, opmar$\Delta_{5,6}$	union	no	30	0.05	0.0093	0.8306	0.1569

Table 3.10.: Brute force results for forecasting one-year sales growth. The three best sets of four, five and six reference variable, respectively, are used in the backtest of PCA rank deviation.

Choice	Ref.Var.	Comb.	Cor.	w	Size	Δ_q	KS	CvM
Best	β, P/E	union	yes	5	0.05	0.0158	2.7215	1.8060
	β, P/E, P/B	union	no	5	0.05	0.0199	2.4706	1.5982
	P/E, P/B	union	no	5	0.05	0.0251	3.0271	2.8965
Best 4	at, β, P/E, P/B	union	no	5	0.05	0.0306	4.3259	4.4872
	seq, β, P/E, P/B	union	no	5	0.05	0.0316	4.1763	4.4604
	at, seq, β, P/E	union	no	5	0.05	0.0320	5.0380	6.2068
Best 5	at, seq, β, P/E, P/B	union	no	5	0.05	0.0323	4.6028	5.0844
	sales, at, seq, P/E, P/B	union	no	5	0.05	0.0374	5.0118	6.1482
	sales, at, β, P/E, P/B	union	yes	5	0.01	0.0379	5.4581	7.0826
Best 6	sales, at, seq, β P/E, P/B	union	no	5	0.05	0.0377	4.9120	6.0955
	sales, opmar, at, β, P/E, P/B	union	yes	5	0.01	0.0437	5.2103	7.8767
	sales, opmar, at, seq, P/E, P/B	union	yes	5	0.01	0.0447	5.2960	8.0278

Table 3.11.: Forward selection results for forecasting three-year sales growth. The three best sets of four, five and six reference variable, respectively, are used in the backtest of PCA rank deviation.

Choice	Ref.Var.	Comb.	Cor.	w	Size	Δ_{q}	KS	CvM
Best single	opmarΔ_6	–	–	30	0.025	0.0157	1.8644	0.8256
	opmarΔ_8	–	–	30	0.025	0.0296	2.0007	1.0991
	opmarΔ_{10}	–	–	30	0.01	0.0319	2.0996	1.1686
	opmarΔ_7	–	–	30	0.01	0.0321	3.4675	3.4485
	opmarΔ_9	–	–	30	0.01	0.0348	2.1389	1.1537
	opmarΔ_5	–	–	30	0.01	0.0363	5.5770	8.9425
Best 2	salesGR$_{10}$, opmarΔ_8	union	yes	30	0.01	0.0239	1.4898	0.5776
	opmar, opmarΔ_7	inters.	no	30	0.05	0.0242	1.8007	0.7059
	opmar$\Delta_{8,10}$	union	no	30	0.05	0.0260	2.0310	1.0684
Best 3	salesGR$_{10}$, opmar$\Delta_{8,9}$	union	yes	30	0.01	0.0223	1.7135	0.7237
	salesGR$_{10}$, opmar$\Delta_{8,10}$	union	yes	30	0.01	0.0255	1.7001	0.6329
	opmar$\Delta_{8:10}$	union	no	30	0.05	0.0261	1.9815	0.9914
Best 4	salesGR$_{10}$, opmar$\Delta_{8:10}$	union	yes	30	0.01	0.0233	1.7000	0.6886
	salesGR$_9$, opmar$\Delta_{8:10}$	union	yes	30	0.01	0.0257	1.8971	0.7147
	salesGR$_8$, opmar$\Delta_{8:10}$	union	no	30	0.01	0.0269	1.7087	0.7064
Best 5	salesGR$_{9,10}$, opmar$\Delta_{8:10}$	union	yes	30	0.01	0.0268	1.6305	0.6640
	salesGR$_{8,9}$, opmar$\Delta_{8:10}$	union	yes	30	0.01	0.0279	1.7261	0.6208
	salesGR$_{8,10}$, opmar$\Delta_{8:10}$	union	yes	30	0.025	0.0283	1.6882	0.7032
Best 6	salesGR$_{8:10}$, opmar$\Delta_{8:10}$	union	yes	30	0.01	0.0293	1.7441	0.8076
	salesGR$_{7,9,10}$, opmar$\Delta_{8:10}$	union	yes	30	0.01	0.0299	2.5688	1.7311
	salesGR$_{7,8,10}$, opmar$\Delta_{8:10}$	union	yes	30	0.01	0.0306	2.5785	1.8136

Table 3.12.: Brute force results for forecasting three-year sales growth. The three best sets of four, five and six reference variable, respectively, are used in the backtest of PCA rank deviation.

Choice	Ref.Var.	Comb.	Cor.	w	Size	Δ_q	KS	CvM
Best	β, P/E, P/B	LARD	–	10	0.05	0.0319	4.7188	6.6191
	β, P/E	LARD	–	30	0.05	0.0388	4.5036	6.1123
	opmar, β	union	yes	30	0.01	0.0392	4.7713	7.8706
Best 4	at, β, P/E, P/B	union	no	5	0.05	0.0594	9.4340	23.8707
	sales, β, P/E, P/B	union	no	5	0.05	0.0655	9.1517	21.6102
	sales, opmar, β, P/E	union	yes	30	0.01	0.0670	7.0775	18.7256
Best 5	at, seq, β, P/E, P/B	union	no	5	0.05	0.0680	9.9804	28.7600
	sales, at, β, P/E, P/B	union	no	5	0.05	0.0688	10.1119	27.1444
	sales, opmar, β, P/E, P/B	union	no	5	0.05	0.0738	9.3260	24.5716
Best 6	sales, at, seq, β, P/E, P/B	union	no	5	0.05	0.0753	10.4470	30.4821
	sales, opmar, at, β, P/E, P/B	union	no	5	0.05	0.0769	10.2567	29.2024
	sales, opmar, seq, β, P/E, P/B	union	yes	30	0.01	0.0816	8.8320	28.1591

Table 3.13.: Forward selection results for forecasting five-year sales growth. The three best sets of four, five and six reference variable, respectively, are used in the backtest of PCA rank deviation.

Choice	Ref.Var.	Comb.	Cor.	w	Size	Δ_{q}	KS	CvM
Best single	opmarΔ_6	–	–	30	0.025	0.0157	1.8644	0.8256
	opmarΔ_{10}	–	–	30	0.025	0.0337	1.7525	0.9926
	opmarΔ_5	–	–	30	0.01	0.0371	3.8690	4.0986
	opmarΔ_9	–	–	30	0.01	0.0375	2.0446	1.4290
	opmarΔ_7	–	–	30	0.01	0.0409	2.5724	1.7135
	opmarΔ_4	–	–	30	0.01	0.0417	4.9117	7.5028
Best 2	salesGR$_{10}$, opmarΔ_6	union	yes	30	0.01	0.0230	2.2279	1.0487
	salesGR$_9$, opmarΔ_6	union	yes	30	0.01	0.0271	1.9119	0.9894
	salesGR$_{10}$, opmarΔ_{10}	union	yes	30	0.01	0.0272	1.7697	1.0131
Best 3	salesGR$_{6,10}$, opmarΔ_6	union	yes	30	0.01	0.0264	1.5280	0.6581
	salesGR$_{8,10}$, opmarΔ_6	union	yes	30	0.01	0.0265	1.7605	0.7962
	salesGR$_{7,10}$, opmarΔ_6	union	yes	30	0.01	0.0275	1.5368	0.6862
Best 4	salesGR$_{5,8,10}$, opmarΔ_6	union	yes	30	0.01	0.0265	2.3651	1.4054
	salesGR$_{5,6,10}$, opmarΔ_6	union	yes	30	0.05	0.0267	2.4667	1.6128
	salesGR$_{5,7,10}$, opmarΔ_6	union	yes	30	0.01	0.0267	2.5496	1.5178
Best 5	salesGR$_{5:7,10}$, opmarΔ_6	union	yes	30	0.05	0.0271	2.3006	1.3987
	salesGR$_{5,6,8,10}$, opmarΔ_6	union	yes	30	0.025	0.0271	2.1439	1.2915
	salesGR$_{5,6,9,10}$, opmarΔ_6	union	yes	30	0.01	0.0279	2.1598	1.2389
Best 6	salesGR$_{5,6,9,10}$, opmar$\Delta_{6,8}$	union	yes	30	0.01	0.0275	2.5795	1.5061
	salesGR$_{5:7,9,10}$, opmarΔ_6	union	yes	30	0.01	0.0277	1.9939	1.2312
	salesGR$_{5:7,10}$, opmar$\Delta_{6,8}$	union	yes	30	0.01	0.0279	2.5866	1.6759

Table 3.14.: Brute force results for forecasting five-year sales growth. The three best sets of four, five and six reference variable, respectively, are used in the backtest of PCA rank deviation.

Choice	Ref.Var.	Comb.	Cor.	w	Size	Δ_q	KS	CvM
Best	β, P/E	LARD	–	30	0.05	0.0492	5.1695	6.0082
	opmar, β	union	yes	30	0.01	0.0525	5.7728	10.7982
	opmar, seq, P/B	inters.	no	30	0.025	0.0543	0.6534	0.0646
Best 4	sales, opmar, β, P/E	union	yes	30	0.025	0.0781	7.6348	21.5909
	sales, opmar, seq, P/E	union	yes	30	0.01	0.0864	8.0867	26.1398
	sales, opmar, β, P/B	union	yes	30	0.01	0.0871	9.0112	25.3800
Best 5	sales, opmar, β, P/E, P/B	union	yes	30	0.01	0.0894	8.8936	25.6606
	sales, opmar, at, β, P/E	union	yes	30	0.01	0.0931	8.5649	29.8700
	sales, opmar, seq, β, P/E	union	yes	30	0.01	0.0943	8.2488	29.5979
Best 6	sales, opmar, seq, β, P/E, P/B	union	yes	30	0.01	0.0975	9.0194	30.8282
	sales, opmar, at, β, P/E, P/B	union	yes	30	0.01	0.1000	9.3992	31.4917
	sales, opmar, at, seq, β, P/E	union	yes	30	0.01	0.1045	9.0507	36.2843

Table 3.15.: Forward selection results for forecasting 10-year sales growth. The three best sets of four, five and six reference variable, respectively, are used in the backtest of PCA rank deviation.

Choice	Ref.Var.	Comb.	Cor.	w	Size	Δ_q	KS	CvM
Best single	opmarΔ_6	–	–	30	0.025	0.0157	1.8644	0.8256
	opmarΔ_7	–	–	30	0.025	0.0461	3.5950	3.8367
	opmarΔ_5	–	–	30	0.025	0.0482	4.1254	5.0874
	opmarΔ_4	–	–	30	0.01	0.0510	4.7700	6.4086
	opmarΔ_8	–	–	30	0.025	0.0520	3.8959	4.4523
	opmarΔ_9	–	–	30	0.05	0.0585	3.8342	4.8646
Best 2	salesGR$_6$, opmarΔ_7	union	yes	30	0.01	0.0401	3.1193	3.2598
	opmar$\Delta_{5,6}$	union	yes	30	0.05	0.0415	4.1217	4.9332
	salesGR$_5$, opmarΔ_7	union	yes	30	0.01	0.0424	3.2929	3.3610
Best 3	salesGR$_6$, opmar$\Delta_{5,7}$	union	yes	30	0.05	0.0409	3.5038	3.7480
	salesGR$_7$, opmar$\Delta_{5,6}$	union	yes	30	0.05	0.0410	3.5776	3.9702
	salesGR$_6$, opmar$\Delta_{5,6}$	union	yes	30	0.01	0.0416	3.6444	3.8220
Best 4	salesGR$_{6,7}$, opmar$\Delta_{5,7}$	union	yes	30	0.01	0.0388	3.0820	3.2019
	salesGR$_7$, opmar$\Delta_{5:7}$	union	yes	30	0.05	0.0411	3.4968	3.5954
	salesGR$_{5,6}$, opmar$\Delta_{5,6}$	union	yes	30	0.01	0.0423	3.4630	3.6062
Best 5	salesGR$_{5:7}$, opmar$\Delta_{5,7}$	union	yes	30	0.025	0.0419	3.2042	3.3900
	salesGR$_{6:8}$, opmar$\Delta_{5,7}$	union	yes	30	0.01	0.0422	3.0932	3.3684
	salesGR$_{5,6}$, opmar$\Delta_{4:6}$	union	yes	30	0.01	0.0425	3.7850	4.3952
Best 6	salesGR$_{6:8}$, opmar$\Delta_{5,7,8}$	union	yes	30	0.01	0.0403	2.9479	2.9308
	salesGR$_{5:7}$, opmar$\Delta_{5:7}$	union	yes	30	0.025	0.0420	3.4675	3.5151
	salesGR$_{6:9}$, opmar$\Delta_{5,7}$	union	yes	30	0.01	0.0432	3.0288	3.4022

Table 3.16.: Brute force results for forecasting 10-year sales growth. The three best sets of four, five and six reference variable, respectively, are used in the backtest of PCA rank deviation.

Choice	Ref.Var.	Comb.	Cor.	w	Size	Δ_q	KS	CvM
Best	β, P/E	LARD	–	30	0.025	0.0680	4.1868	4.2020
	opmar, P/E	union	yes	30	0.01	0.0698	4.5438	7.8400
	at, β, P/E, P/B	LARD	–	30	0.05	0.0774	6.1698	10.9065
Best 4	opmar, β, P/E, P/B	LARD	–	20	0.025	0.0804	9.2137	19.5756
	opmar, at, β, P/E	LARD	–	20	0.05	0.0941	11.6692	31.5442
	sales, β, P/E, P/B	LARD	–	30	0.05	0.0956	7.3187	16.8942
Best 5	opmar, at, β, P/E, P/B	LARD	–	20	0.05	0.0989	8.7529	21.4948
	sales, opmar, at, seq, β	inters.	yes	30	0.05	0.1059	4.1239	6.7161
	sales, opmar, β, P/E, P/B	union	yes	30	0.01	0.1217	8.2634	22.2247
Best 6	sales, opmar, at, β, P/E, P/B	union	yes	30	0.01	0.1389	9.1815	26.8634
	sales, opmar, seq, β, P/E, P/B	union	yes	30	0.025	0.1414	9.7186	28.1473
	sales, opmar, at, seq, P/E, P/B	union	yes	30	0.01	0.1443	9.2700	28.0870

Bibliography

W. Antweiler. Pacific exchange rate service. Sauder School of Business, University of British Columbia, 2015. URL `http://fx.sauder.ubc.ca/`.

A.K. Anundsen. Econometric regime shifts and the US subprime bubble. *Journal of Applied Econometrics*, 30(1):145–169, 2015.

J.S. Armstrong. The forecasting canon: Nine generalizations to improve forecast accuracy. *International Journal of Applied Forecasting*, 1:29–35, 2005.

A. Arsova and D.D.K. Örsal. A panel cointegrating rank test with structural breaks and cross-sectional dependence. *Econometrics and Statistics*, 17:107–129, 2021.

R.H. Ashton and A.M. Cianci. Motivational and cognitive determinants of buy-side and sell-side analyst earnings forecasts: An experimental study. *Journal of Behavioral Finance*, 8(1):9–19, 2007.

Bank of England. United Kingdom: Deposit interest rate. The Global Economy, 2020. URL `https://www.theglobaleconomy.com/United-Kingdom/deposit_interest_rate/`.

J. Batselier and M. Vanhoucke. Practical application and empirical evaluation of reference class forecasting for project management. *Project Management Journal*, 47(5):36–51, 2016.

J. Berkson. Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association*, 33(203):526–536, 1938.

S. Bhojraj and C.M.C. Lee. Who is my peer? A valuation-based approach to the selection of comparable firms. *Journal of Accounting Research*, 40(2):407–439, 2002.

Bitcoincharts. Historic trade data. Bitcoincharts, 2017. URL `http://api.bitcoincharts.com/v1/csv/`.

Board of Governors of the Federal Reserve System. 1-month eurodollar deposit rate London. FRED, Federal Reserve Bank of St. Louis, 2020. URL `https://fred.stlouisfed.org/series/DED1/`.

P. Bordalo, N. Gennaiolo, R. la Porta, and A. Shleifer. Diagnostic expectations and stock returns. *The Journal of Finance*, 74(6):2839–2874, 2019.

R.F. Bordley. Reference class forecasting: Resolving its challenge to statistical modeling. *The American Statistician*, 68(4):221–229, 2014.

J. Breitung and M. Pesaran. Unit roots and cointegration in panels. In L. Matyas and P. Sevestre, editors, *The Econometrics of Panel Data: Fundamentals and Recent Developments in Theory and Practice*, chapter 9, pages 279–322. Springer, 2008.

J.B. Bremnes. Probabilistic forecasts of precipitation in terms of quantiles using nwp model output. *Monthly Weather Review*, 132(1):338–347, 2004.

T.S. Breusch and A.R. Pagan. The lagrange multiplier test and its applications to model specification in econometrics. *The Review of Economic Studies*, 47(1):239–253, 1980.

D. Bunn and G. Wright. Interaction of judgemental and statistical forecasting methods: Issues and analysis. *Management Science*, 37(5):501–518, 1991.

K.C. Butler and H. Saraoglu. Improving analysts' negative earnings forecasts. *Financial Analysts Journal*, 55(3):48–56, 1999.

Center for Research in Security Prices. CRSP daily stock. Wharton Research Data Services, 2020. URL `https://wrds-www.wharton.upenn.edu/`.

L.K. Chan, J. Karceski, and J. Laknoishok. The level and persistence of growth. *The Journal of Finance*, 63(2):643–684, 2003.

W. Chang, E. Chen, B.A. Mellers, and P.E. Tetlock. Developing expert political judgment: The impact of training and practice on judgmental accuracy in geopolitical forecasting tournaments. *Judgment and Decision Making*, 11:509–526, 2016.

E.-T. Cheah and J. Fry. Speculative bubbles in bitcoin markets? An empirical investigation into the fundamental values of bitcoin. *Economics Letters*, 130:32–36, 2015.

E.K. Cheng. A practical solution to the reference class problem. *Columbia Law Review*, 109(8):2081–2106, 2009.

C.-S. J. Chu, M. Stinchcombe, and H. White. Monitoring structural change. *Econometrica*, 64(5):1045–1065, 1996.

A.C. Cooper, C.Y. Woo, and W.C. Dunkelberg. Entrepreneurs' perceived chances for success. *Journal of Business Venturing*, 3(2):97–108, 1988.

S. Corbet, B. Lucey, and L. Yarovaya. Datestamping the bitcoin and ethereum bubbles. *Finance Research Letters*, 26:81–88, 2018.

A. Cretarola and G. Figà-Talamanca. Detecting bubbles in bitcoin price dynamics via *Market Exuberance. Annals of Operations Research*, 299:459–479, 2021.

A.P. Dawid. Statistical theory: The prequential approach. *Journal of the Royal Statistical Society. Series A (General)*, 147(2):278–292, 1984.

B. De Bruijn and P.H. Franses. Heterogeneous forecast adjustment. *Journal of Forecasting*, 36(4):337–344, 2017.

C. Decker and R. Wattenhofer. Bitcoin transaction malleability and mtgox. In M. Kutyłowski and J. Vaidya, editors, *Computer Security - ESORICS 2014*, pages 313–326. Springer International Publishing, 2014.

H. Dette and D. Wied. Detecting relevant differences in time series models. *Journal of the Royal Statistical Society, Series B*, 78(2):371–394, 2016.

F. X. Diebold, T. A. Gunther, and A. S. Tay. Evaluating density forecasts with applications to financial risk management. *International Economic Review*, 39(4):863–883, 1998.

F.X. Diebold. *"Big Data" Dynamic Factor Models for Macroeconomic Measurement and Forecasting: A Discussion of the Papers by Lucrezia Reichlin and by Mark W. Watson*, pages 115–122. Econometric Society Monographs. Cambridge University Press, 2003.

H. Dong and W. Dong. Bitcoin: Exchange rate parity, risk premium, and arbitrage stickiness. *British Journal of Economics, Management & Trade*, 5(1):105–113, 2014.

N. Du and D.V. Budescu. How (over) confident are financial analysts? *Journal of Behavioral Finance*, 19(3):308–318, 2018.

N. Du and J.E. McEnroe. Are multiple analyst earnings forecasts better than the single forecast? *Journal of Behavioral Finance*, 12(1):1–8, 2011.

European Money Markets Institute. Euribor. Triami Media B.V., 2020. URL `https://www.euribor-rates.eu/de/euribor-werte-pro-jahr/`.

FactSet. Core company data - estimates data. FactSet, 2021. URL `http://factset.com/`.

W. Fan and A. Bifet. Mining big data: Current status, and forecast to the future. *Special Interest Group on Knowledge Discovery in Data Explorations Newsletter*, 14(2):1–5, 2013.

C. Fink and T. Johann. Bitcoin markets. *SSRN Working Paper (No. 2408396)*, 2014.

B. Flyvbjerg. From nobel prize to project management: Getting risks right. *Project Management Journal*, 37(3):5–15, 2006.

B. Flyvbjerg. Curbing optimism bias and strategic misrepresentation in planning: Reference class forecasting in practice. *European Planning Studies*, 16(1):3–21, 2008.

R. Gibrat. *Les Inegalites economiques*. Recueil Sirey, 1931.

I. Gilboa, O. Lieberman, and D. Schmeidler. Empirical similarity. *The Review of Economics and Statistics*, 88(3):433–444, 2006.

T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

T. Gneiting, F. Balabdaoui, and A. E. Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B*, 69(2):243–268, 2007.

K.C. Green and J.S. Armstrong. Structured analogies for forecasting. *International Journal of Forecasting*, 23(3):365–376, 2007.

J.B. Guerard, H. Markowitz, and G. Xu. Earnings forecasting in a global stock selection model and efficient portfolio construction and management. *International Journal of Forecasting*, 31(2):550–560, 2015.

T. M. Hamill. Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, 129(3):550 – 560, 2001.

L. Held, B. Schrödle, and H. Rue. *Posterior and Cross-validatory Predictive Checks: A Comparison of MCMC and INLA*, pages 91–110. Physica-Verlag HD, Heidelberg, 2010.

R. Hull. Credit ratings and firm value. *Investment Management and Financial Innovations*, 17:157–168, 6 2020.

R. Jame, R. Johnston, S. Markov, and M.C. Wolfe. The value of crowdsourced earnings forecasts. *Journal of Accounting Research*, 54(4):1077–1110, 2016.

M. Jansson. Consistent covariance matrix estimation for linear processes. *Econometric Theory*, 18(6):1449–1459, 2002.

I. T. Jolliffe. *Principal Component Analysis*. Springer New York, NY, 2nd edition, 2002.

S. Jones and D. Johnstone. Analyst recommendations, earnings forecasts and corporate bankruptcy: Recent evidence. *Journal of Behavioral Finance*, 13(4):281–298, 2012.

D. Kahneman and D. Lovallo. Timid choices and bold forecasts: A cognitive perspective on risk taking. *Management Science*, 39(1):17–31, 1993.

D. Kahneman and A. Tversky. On the psychology of prediction. *Psychological Review*, 80(4):237–251, 1973.

D. Kahneman and A. Tversky. Intuitive prediction: Biases and corrective procedures. *Management Science*, 12:313–327, 1979.

C.W. Karvetski, C. Meinel, D.T. Maxwell, Y. Lu, B.A. Mellers, and P.E. Tetlock. What do forecasting rationales reveal about thinking patterns of top geopolitical forecasters? *International Journal of Forecasting*, 2021.

D.M. Kent, E. Steyerberg, and D. van Klaveren. Personalized evidence based medicine: Predictive approaches to heterogeneous treatment effects. *British Medical Journal*, 363, 2018.

R. Kitchin and G. McArdle. What makes big data, big data? Exploring the ontological characteristics of 26 datasets. *Big Data & Society*, 3(1), 2016.

J.O. Knudsen, S. Kold, and T. Plenborg. Stick to the fundamentals and discover your peers. *Financial Analysts Journal*, 73(3):85–105, 2017.

T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43:59–69, 1982.

S. Kunte. The herding mentality: Behavioral finance and investor biases. https://blogs.cfainstitute.org/investor/2015/08/06/the-herding-mentality-behavioral-finance-and-investor-biases/, 2015.

B. Lee, J. O'Brien, and K. Sivaramakrishnan. An analysis of financial analysts' optimism in long-term growth forecasts. *Journal of Behavioral Finance*, 9(3):171–184, 2008.

G. Löffler. Biases in analyst forecasts: cognitive, strategic or second-best? *International Journal of Forecasting*, 14(2):261–275, 1998.

T. Lim. Rationality and analysts' forecast bias. *The Journal of Finance*, 56(1):369–385, 2001.

P.S. Lintilhac and A. Tourin. Model-based pairs trading in the bitcoin markets. *Quantitative Finance*, 17(5):703–716, 2017.

G.J. Lobo. Alternative methods of combining security analysts' and statistical forecasts of annual corporate earnings. *International Journal of Forecasting*, 7(1):57–63, 1991.

D. Lovallo and D. Kahneman. Delusions of success: How optimism undermines executives' decisions. *Harvard Business Review*, 81(7):56–63, 117, 2003.

D. Lovallo, C. Clarke, and C. Camerer. Robust analogizing and the outside view: Two empirical tests of case-based decision making. *Strategic Management Journal*, 33: 496–512, 2012.

G.L. Marcialis and F. Roli. Fusion of appearance-based face recognition algorithms. *Pattern Analysis and Applications*, 7:151–163, 2004.

M. Marozzi. A method to address the effectiveness of the sic code for selecting comparable firms. *Electronic Journal of Applied Statistical Analysis*, 6(2):186–201, 2011.

M.J. Mauboussin and D. Callahan. *The Base Rate Book – Sales Growth.* Credit Suisse: Global Financial Strategies, 2015.

H.R. Moon. A note on fully-modified estimation of seemingly unrelated regressions models with integrated regressors. *Economics Letters*, 65:25–31, 1999.

A. Natarajan. Reference class forecasting and machine learning for improved offshore oil and gas megaproject planning: Methods and application. *Project Management Journal*, 53(5):456–484, 2022.

J.Y. Park and M. Ogaki. Seemingly unrelated canonical cointegrating regressions. *Rochester Center for Economic Research: Working Paper No. 280*, 1991.

P. Parveen and B. Thuraisingham. Face recognition using multiple classifiers. In *18th IEEE International Conference on Tools with Artificial Intelligence*, pages 179–186, 2006.

P. Pasquariello. Financial market dislocations. *Review of Financial Studies*, 27(6): 1868–1914, 2014.

P.C.B. Phillips and S. Durlauf. Multiple time series regression with integrated processes. *Review of Economic Studies*, 53:473–495, 1986.

P.C.B. Phillips and B.E. Hansen. Statistical inference in instrumental variables regression with I(1) processes. *Review of Economic Studies*, 57(1):99–125, 1990.

P.C.B. Phillips and H.R. Moon. Linear regression limit theory for nonstationary panel data. *Econometrica*, 67(5):1057–1111, 1999.

M.E. Porter. How competitive forces shape strategy. *Harvard Business Review*, 57(2): 137–145, 1979.

M.E. Porter. *The Competitive Advantage: Creating and Sustaining Superior Performance*. Free Press, 1985.

S. Ramnath, S. Rock, and P. Shane. Value line and I/B/E/S earnings forecasts. *International Journal of Forecasting*, 21(1):185–198, 2005.

H. Reichenbach. *The Theory of Probability: An Inquiry into the Logical and Mathematical Foundations of the Calculus of Probability*. University of California Press, 1949.

J. Reynolds, L. Sögner, M. Wagner, and D. Wied. Deviations from triangular arbitrage parity in foreign exchange and bitcoin markets. *TU Dortmund SFB 823 Working Paper 09/18*, 2018.

J. Reynolds, L. Sögner, and M. Wagner. Deviations from triangular arbitrage parity in foreign exchange and bitcoin markets. *Central European Journal of Economic Modelling and Econometrics*, 12(2):105–146, 2021.

N.R. Sanders and L.P. Ritzman. Judgmental adjustment of statistical forecasts. In J. Scott Armstrong, editor, *Principles of Forecasting: A Handbook for Researchers and Practitioners*, pages 405–416. Springer US, Boston, MA, 2001.

T. Servranckx, M. Vanhoucke, and T. Aouam. Practical application of reference class forecasting for cost and time estimations: Identifying the properties of similarity. *European Journal of Operational Research*, 295(3):1161–1179, 2021.

Y. Shin. A residual-based test of the null of cointegration against the alternative of no cointegration. *Econometric Theory*, 10(1):91–115, 1994.

O. Shmueli, N. Pliskin, and L. Fink. Can the outside-view approach improve planning decisions in software development projects? *Information Systems Journal*, 26(4): 395–418, jul 2016.

S&P Global Market Intelligence. Compustat daily updates - fundamentals annual. Wharton Research Data Services, 2020. URL `https://wrds-www.wharton.upenn.edu/`.

M.H.R. Stanley, L.A.N. Amaral, S.V. Buldyrev, S. Havlin, H. Leschhorn, P. Maass, M.A. Sallinger, and H.E. Stanley. Scaling beahviour in the growth of companies. *Nature*, 379:804–806, 1996.

Statistics Sweden. Sweden: Deposit interest rate. The Global Economy, 2020. URL `https://www.theglobaleconomy.com/Sweden/deposit_interest_rate/`.

O. Stotz and R. von Nitzsch. The perception of control and the level of overconfidence: Evidence from analyst earnings estimates and price targets. *Journal of Behavioral Finance*, 6(3):121–128, 2005.

P.E. Tetlock and D. Gardner. *Superforecasting: The Art and Science of Prediction.* Random House, 2016.

E. Theising. *Monitoring Cointegration in a System of Homogeneous Cointegrating Regressions.* Unpublished Master Thesis, 2018.

E. Theising. Distributional reference class forecasting of corporate sales growth with multiple reference variables. *working paper, arXiv:2405.03402*, 2024.

E. Theising and D. Wied. Monitoring cointegration in systems of cointegrating relationships. *Econometrics and Statistics*, forthcoming, 2023.

E. Theising, D. Wied, and D. Ziggel. Reference class selection in similarity-based forecasting of corporate sales growth. *Journal of Forecasting*, 42(5):1069–1085, 2023.

T.N. Themsen. The processes of public megaproject cost estimation: The inaccuracy of reference class forecasting. *Financial Accountability & Management*, 35(4):337–352, 2019.

A. Tversky and D. Kahneman. Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, 5(2):207–232, 1973.

A. Tversky and D. Kahneman. Judgement under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131, 1974.

U.S. Bureau of Labor Statistics. Consumer price index for all urban consumers: All items in US city average. FRED, Federal Reserve Bank of St. Louis, 2020. URL `https://fred.stlouisfed.org/series/CPIAUCSL/`.

J. Venn. *The Logic of Chance*. MacMillan and Co., 3rd edition, 1888.

T.J. Vogelsang and M. Wagner. Integrated modified OLS estimation and fixed-*b* inference for cointegrating regressions. *Journal of Econometrics*, 178(2):741–760, 2014.

M. Wagner and D. Wied. Consistent monitoring of cointegrating relationships: The US housing market and the subprime crisis. *Journal of Time Series Analysis*, 38(6): 960–980, 2017.

C. Wolfe and B. Flores. Judgmental adjustment of earnings forecasts. *Journal of Forecasting*, 9(4):389–405, 1990.

World Bank. Deposit interest rate. The World Bank Group, 2020. URL `https://data.worldbank.org/indicator/FR.INR.DPST/`.

G. Yu and J. Zhang. Revisit capital control policies when bitcoin is in town. *SSRN Working Paper (No. 3053474)*, 2017.

V. Zarikas and C.P. Kitsos. Risk analysis with reference class forecasting adopting tolerance regions. In C.P. Kitsos, T.A. Oliveira, A. Rigas, and S Gulati, editors, *Theory and Practice of Risk Assessment*, pages 235–247. Springer International Publishing, 2015.

Appendix A.

Software and Data

All computations in this dissertation were implemented in R (https://www.r-project.org/). The simulations and backtests were parallelized and performed using CHEOPS, a scientific High Performance Computer at the Regional Computing Center of the University of Cologne funded by the DFG (Funding number: INST 216/512/1FUGG).

The data that support the findings in Chapter 1 of this dissertation are available from the following sources under stated restrictions. The Bitcoin data are openly available from Bitcoincharts at https://bitcoincharts.com/ and were downloaded on 16 October 2017. Restrictions apply to the use of these data to non commercial purpose only. The exchange rate data of fiat currencies are openly available from the Pacific Exchange Rate Service at http://fx.sauder.ubc.ca/. Restrictions apply to the use of these data to research purpose only. The Eurodollar deposit rates data are openly available from the Federal Reserve Bank of St. Louis at https://fred.stlouisfed.org/, reference DED1. The Euribor rate data are openly available at https://www.euribor-rates.eu/ and were downloaded on 30 September 2020. Restrictions apply to the use of these data to non commercial purpose only. The GBP and SEK deposit interest rate data are available from TheGlobalEconomy.com and were downloaded on 30 September 2020. Restrictions apply to the availability of these data, which were used under license for this dissertation. Data are available at https://www.theglobaleconomy.com/ with the permission of TheGlobalEconomy.com. The AUD, CAD and RUB deposit interest rate data are openly available from the World Bank at https://data.worldbank.org/ and were downloaded on 30 September, 2020. The data were used under a “Creative Commons Attribution 4.0 International License”.

Moreover, the data that support the findings in Chapters 2 and 3 of this dissertation are available from the following sources under stated restrictions. Wharton Research Data Services (WRDS) was used in preparing these chapters. This service and the data available thereon constitute valuable intellectual property and trade secrets of

WRDS and/or its third-party suppliers. The “CRSP daily stock” and “Compustat daily updates - fundamentals annual” data are available from WRDS and were downloaded on 28 and 30 January, 2020, respectively. Restrictions apply to the availability of these data, which were used under license for this dissertation. Data are available at https://wrds-www.wharton.upenn.edu/ with the permission of WRDS. The consumer price index data are openly available at FRED (Federal Reserve Economic Data) at https://fred.stlouisfed.org/, reference CPIAUCSL, and were downloaded on 29 January, 2020. The “Core company data - estimates data” were provided by colleagues at Flossbach von Storch AG, are available from FactSet and were downloaded on 7 January, 2021. Restrictions apply to the availability of these data, which were used under license for this dissertation. Data are available at http://factset.com/ with the permission of FactSet.