Erasmus University Rotterdam
Erasmus School of Economics
Master Thesis Econometrics and Management Science

Assessing the Benefits of Save Anytime-Valid Inference for Macroeconomic Data

Yonis Kulane (545931)

Supervisor:	dr. Nick W. Koning
Second assessor:	dr. Andreas Pick
Date final version:	1st October 2023

Abstract

The confidence sequences derived by Choe & Ramdas (2022) to compare the forecast quality of two forecasters are anytime-valid but come at a cost: loss of power. We assess the loss of power for sample sizes smaller than 600 to judge whether confidence sequence can be applied to macroeconomic data. We perform a simulation study to compare the power of confidence sequences against that of the DM and GW tests. We also assess their anytime-validity. Confidence sequences need more than twice as much data to achieve the same power as the DM test. In addition, confidence sequences need more than 100 observations more than the DM test to detect a change in forecaster performance. This amounts to 25 years of quarterly or roughly eight years of monthly data. Finally, we implement confidence sequences to compare directional forecasts constructed from consumer sentiment, professional forecasters, and economists' expectations. We find that confidence sequences can detect past changes in forecaster performance.

Contents

1 Introduction

The Diebold & Mariano (1995) test (DM) is widely used to assess if forecaster A is significantly better than forecaster B. Aside from the stationarity assumption, it suffers from an inflated type I error rate in the case of continuous monitoring. Anytime-valid approaches, such as the confidence sequences developed by Choe & Ramdas (2022), offer a valid alternative under continuous monitoring. However, this anytime-valid property comes at a cost: loss in power.

Nevertheless, Choe & Ramdas (2022) states that anytime-valid methods do not need larger sample sizes than the DM and Giacomini & White (2006) (GW) tests for high power. They base their results on a simulation study involving a sample size of 10,000 observations. Such a large amount of data is unrealistic for macroeconomic data such as GDP and inflation. Their sample size corresponds to 833 and 2500 years of monthly and quarterly observations, respectively. In contrast, we limit ourselves to 600 observations, which corresponds to 50 and 150 years of monthly and quarterly observations, respectively.

Such a smaller sample size ensures that our results are generalizable to macroeconomic data. This aids, among others, policymakers who could utilise anytime-valid methods to select the best forecaster. Accurate forecasts are needed to support decision-making, and identifying the best forecaster is therefore essential. However, the best forecaster can vary over time. Anytime-valid methods might demonstrate how this finding might differ over time as new data becomes available. Macroeconomic data are published at a low frequency: mainly annually, quarterly or monthly. Having a method that can incorporate the latest information and allow for inference considerably aids policymakers, in particular, who rely on these forecasts.

In this research, we investigate the advantages and disadvantages of confidence sequences (CS) relative to the Diebold-Mariano (DM) and Giacomini-White (GW) tests for evaluating probability forecasts in cases with small sample sizes, $T \leq 600$. We assess the power of confidence sequences and the DM and GW tests, and their type I errors under continuous monitoring through a simulation study.

We perform a power analysis to study the relationship between the effect size, sample size, and power of CS and the DM and GW tests. Moreover, we examine how much data CS, DM and GW tests need to detect a change in forecaster performance. In addition, we examine the (cumulative) type I error to contrast their anytime-validity. Lastly, we present

an empirical application of confidence sequences to directional forecasts constructed from consumer, economist and professional forecaster expectations.

We find that the loss in power of confidence sequences is significant. Confidence sequences need more than twice as much data as the DM test to reject the null hypothesis when it does not hold. In the scenario where a trend is introduced in the score differentials, confidence sequences need three times as many or 100 observations more than the DM test to reject the null hypothesis. This corresponds to more than 25 years of quarterly or roughly eight years of monthly macroeconomic data.

As expected, unlike the DM and GW tests, the confidence sequences are anytime-valid. The cumulative type I error of the CS is and remains zero, while it rapidly grows large for the DM test and reaches up to 27%. The GW test performs slightly better in this regard, although it also is not anytime-valid. Its cumulative type I error reaches 20%.

Our empirical application to directional forecasts shows that CS can be used in practice given enough data, and there is a significant gap in forecasting performance between two competing forecasters. We find that forecasts constructed from economist and professional forecaster expectations on average outperform those constructed from consumer sentiment. In addition, CS can detect changes in past forecasters' performance.

These findings demonstrate challenges to using anytime-valid methods to compare macroeconomic forecasts as data comes in slowly, particularly yearly data. These methods are valid under continuous monitoring, but lack sufficient power to quickly detect changes in forecast performance to be useful in live testing. However, they can be used to 'look backwards' and examine whether Forecaster A outperformed Forecaster B on average in the past and how that changed.

This paper proceeds as follows. Section 2 discusses the literature surrounding forecast evaluation of continuous, binary and directional forecasts. Anytime-valid confidence sequences are discussed in Section 3. Our simulation study is presented in Section 4 and in Section 5 we apply confidence sequences to compare directional forecasts. Section 6 concludes the paper with suggestions for future research.

2 Literature Forecast Evaluation & Directional Forecasts

Forecast evaluation has mostly been concerned with accuracy metrics to get a sense of the forecast accuracy. For continuous outcomes, the most popular metric is the mean-squared error (MSE). Other accuracy metrics exist such as the MAE, RMSE and accuracy metrics that are scale-independent. Hewamalage et al. (2023) summarised it as follows. Many different point forecast accuracy measures have been proposed in the forecasting literature based on (i) whether squared or absolute errors are used (ii) techniques used to make them scale-free and (iii) operators such as mean, and median used to summarize the errors.

In this research, we focus on probability forecasts of binary outcomes. These make use of other accuracy metrics. Among those, the Brier score proposed by Brier (1950) is widely used as Lai et al. (2011) states. Examples of other score functions are the logarithmic score proposed by I. J. Good (1952) and the spherical score (I. Good, 1971). These score functions are examples of proper score functions and Choe & Ramdas (2022) state that they are the main approach to evaluate probabilistic forecasts. The reason is that they assess both calibration and sharpness. Calibration is defined as the statistical consistency between the distributional forecasts and the observations, and sharpness refers to the concentration of the predictive distributions (Gneiting et al., 2007).

These measures of accuracy enable us to give different scores to competing models. The natural question that then arises is: Which model performs best? To address this question, several methods have been proposed. Stekler (1991) has proposed three ways in which the statistical significance of the difference in model accuracy can be tested: the MSE regression test developed by Ashley et al. (1980), analysing the percentage of times forecaster A is better than forecaster B and lastly the Wilcoxon (1947) Rank Sum Test. Other tests include the Diebold & Mariano (1995) (DM) test, Giacomini & White (2006) (GW) test, F-test and Friedman test. The latter two are used to compare more than two forecasts. For a comprehensive overview of which test is applicable in different situations, refer to Hewamalage et al. (2023) (see Figure 1).

Among these tests, the DM and GW tests can be applied to compare two probability forecasts. The DM test is the first formal test that compared predictive accuracy as

stated by Diebold (2015). The null hypothesis is that the expected loss-differentials of two forecasters are equal to zero. Diebold & Mariano (1995) derived the asymptotic normality of the test statistic. Its only assumption is that the loss-differentials are covariance stationary. On the contrary, Giacomini & White (2006) formulate a test of conditional predictive ability. Unlike the DM test, the GW test allows for nonstationary score differentials. Due to its widespread adoption (Diebold, 2015), the DM test in addition to the GW test will serve as our benchmark against which we examine the performance of anytime-valid methods.

The evaluation metrics mentioned above measure the statistical accuracy of forecasts. In economic decision problems, statistical accuracy does not always conform to economic utility. As Blaskowitz & Herwartz (2014) states: "the squared forecast error provides only a partial assessment of economic forecasts". This has also been noted by Diebold & Mariano (1995) and Granger & Pesaran (2000), among others. In a lot of instances, forecasters are not only interested in the level or size of a forecast but also in the direction of a forecast. Examples of these include an investor who buys a stock if they expect the stock to increase in value or a central bank that raises interest rates if inflation is expected to increase (Blaskowitz & Herwartz, 2011).

Directional accuracy has been used to evaluate the GDP and inflation forecasts. See Tsuchiya (2016) for an overview. Various methods have been employed to assess directional accuracy. These include the exact test of Fisher (1922) based on 2x2 contingency tables, the test proposed by H. Pesaran & Timmermann (1992), and the less frequently used test introduced by M. H. Pesaran & Timmermann (2009). The null hypothesis in these tests states that the direction of change in a forecast and the realisation are independent. A forecast is considered a useful predictor of the direction of change if the null hypothesis is rejected (Tsuchiya, 2013). However, these tests have certain methodological drawbacks.

The aforementioned tests, except for the M. H. Pesaran & Timmermann (2009) test, tend to over-reject the null hypothesis if the forecasts are serially correlated. Additionally, none of these tests directly examine the comparative performance of different forecasters. Rather, they assess the correlation between a forecast and the direction of change.

In contrast, this research evaluates whether anytime-valid confidence sequences that have little distributional assumptions are practical when comparing two directional fore-

casters. We evaluate predictions of binary outcomes, namely directional forecasts, and thus use the Brier score. We contrast the performance of confidence sequences with the performance of the DM and GW tests. We follow Vrontos et al. (2021) and introduce the following binary variable

$$D_t = \begin{cases} 1, & \text{if } Y_t > Y_{t-1} \\ 0, & \text{if } Y_t \leq Y_{t-1}, \end{cases}$$

where D_t is the variable denoting the directional change and Y_t is the underlying time series. This novel approach has not been implemented before in this strand of literature despite its economic relevance.

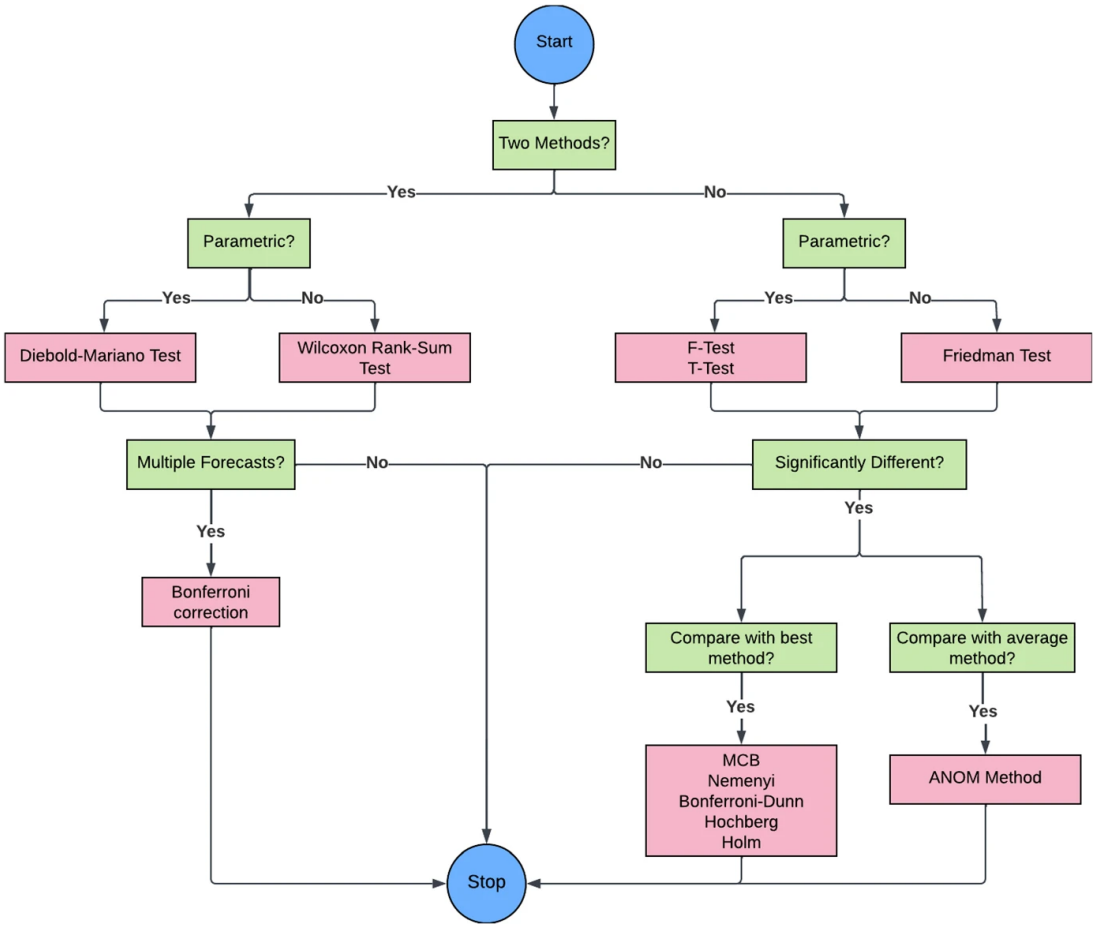


Figure 1: Flow chart from Hewamalage et al. (2023) for statistical tests selection to measure significance of differences between models

3 Anytime-Valid Confidence Sequences

In this subsection, we present the Confidence Sequences we use to compare competing binary forecasts derived by Choe & Ramdas (2022). First, we give a definition of confidence sequences.

Confidence sequences are a sequence of confidence intervals that are uniformly valid over a time horizon (Choe & Ramdas, 2022). A $(1-\alpha)$ confidence sequence (CS) for a time-varying sequence of target parameters $(\theta_t)_{t=1}^{\infty}$ can be defined as a sequence of confidence intervals (CIs) $(C_t)_{t=1}^{\infty}$ such that the probability of any of the CIs excluding θ_t is not greater than α, $P(\exists t \geq 1 : \theta \notin C_t) \leq \alpha$. Howard et al. (2021) shows that this guarantee holds even for arbitrary stopping times. This property is referred to as anytime-valid (Choe & Ramdas, 2022).

In the following, we discuss the literature on confidence sequences and anytime-valid inference. Afterwards, we explain martingales, Ville's inequality, and scoring rules that are used to construct confidence sequences. Finally, we present the confidence sequences and how they are formed. A more detailed derivation and proofs can be found in Choe & Ramdas (2022) and Howard et al. (2021).

3.1 Related Works

Confidence Sequences have their origins in the works of Darling & Robbins (1967); Robbins (1970); Lai (1976) (Choe & Ramdas, 2022). Darling & Robbins (1967) derived confidence sequences for the mean, median and variance of i.i.d. random variables, Robbins (1970), among others, formulated a confidence sequence for the mean of a normal distribution with known variance and the median of i.i.d. random variables, and Lai (1976) used moment generating function martingales to form confidence sequences for the parameters of the binomial, Poisson and gamma distribution.

In the clinical trial literature, confidence sequences are known as repeated confidence intervals (Jennison & Turnbull, 1984, 1989), and other names include always-valid confidence intervals (Johari et al., 2019) and anytime confidence intervals (Jamieson & Jain, 2018), where the latter is used in the machine learning literature (Howard et al., 2021). Renewed attention to confidence sequences has been sparked in part by recent literature on the best-arm identification in multi-armed bandits (Jamieson et al., 2014; Jamieson &

Jain, 2018).

This research builds on the work of Choe & Ramdas (2022) who combined the Empirical Bernstein confidence sequences derived by Howard et al. (2021) that are anytime-valid and have no distributional assumptions with the martingale property of forecast score differentials discussed by Lai et al. (2011) to develop a sequential procedure for forecast evaluation. This research studies the power of Empirical Bernstein confidence sequences for forecast score differentials proposed in small samples and contrasts its power with that of not anytime-valid parametric tests, DM and GW tests.

3.2 Prerequisites

The confidence sequences defined later in this section are based on martingales. These martingales will be bounded using Ville's inequality (Ville, 1939) which is a generalisation of Markov's inequality to martingales. For ease of explanation, we first show how random variables can be bounded by Markov's inequality and how Ville's inequality generalises this idea to martingales. Lastly, we elaborate on scoring rules as the confidence sequences bound the difference in score functions.

Consider a non-negative random variable, $X \geq 0$, and let $\alpha > 0$ be a constant. Markov's inequality states that the probability of X exceeding α is bounded as follows

$$P\left(X \geq \frac{\mathbb{E}[X]}{\alpha}\right) \leq \alpha.$$

Ville's inequality generalises this inequality to martingales. A martingale is a series $(X_t)_{t=0}^{\infty}$ that given a certain probability distribution P and information set $\mathcal{I}$ satisfies the following property:

$$\mathbb{E}[|X_t|] < \infty, \quad \text{for all } t \geq 0, \tag{1}$$

$$\mathbb{E}_t[X_{t+h}] = X_t, \quad \text{for all } t, h \geq 0, \tag{2}$$

where $\mathbb{E}_{t-1}[\cdot] = \mathbb{E}_P[\cdot|\mathcal{I}_{t-1}]$. A martingale is a supermartingale if the equality sign in Equation 2 is a $\leq$ sign and a submartingale if Equation 2 reads $\mathbb{E}_t[X_{t+h}] \geq X_t, \forall t, h \geq 0$. In addition, Shafer et al. (2011) defines a test martingale as a nonnegative martingale starting at one ($X_0 = 1$).

Ville's inequality (Ville, 1939) states that if $(X_t)_{t=1}^{\infty}$ is a test supermartingale under P, then for any $\alpha \in (0, 1)$

$$P(\exists t \geq 1 : X_t > 1/\alpha) \leq \alpha,$$

which holds for every time t and without a predetermined sample size.

The second concept we need is scoring rules. Scoring rules are used to evaluate the quality of a forecaster. The CS bound the difference in score functions of two forecasters using Ville's inequality. In a binary outcome space, $\mathcal{Y} = \{0, 1\}$, with a set of probability forecasts $\mathcal{P} = [0, 1]$, the performance of a (probabilistic) forecast $p \in \mathcal{P}$ given an observation $y \in \mathcal{Y}$ is evaluated using a scoring rule which is a real-valued function, $\mathcal{S} : \mathcal{P} \times \mathcal{Y} \to \mathbb{R}$. The scoring rule used in this research is the Brier score $S(p, y) = 1 - (p - y)^2$ proposed by Brier (1950).

3.3 Game-theoretic setup

The method of Choe & Ramdas (2022) assumes that the forecasts and outcomes are generated as follows. Consider a forecasting game in which two forecasters each make a forecast about an event that occurs over time t. A third participant, reality, generates the outcomes that the forecasters are attempting to predict from a sequence of distributions. At the start of each round, $t = 1, 2, ..,$

1. Forecaster A and Forecaster B construct their forecasts $a_t, b_t \in [0, 1]$.

2. Reality chooses $r_t \in [0, 1]$. The choice of r_t will not be revealed to the forecasters.

3. The outcome is sampled, $y_t \sim \text{Bernoulli}(r_t)$ and is revealed to the forecasters.

This game is observed by an outsider, who compares the performance of the two forecasters only based on the observed data $(a_t, b_t, y_t)_{t=1}^{\infty}$. Note that no distributional assumptions are made about how the forecasts are generated.

3.4 Deriving Confidence Sequences

In this subsection, we present confidence sequences that Choe & Ramdas (2022) constructed using martingales and Ville's inequality. We first discuss the (pointwise) score

differential and the cumulative score differential. Next, we discuss how the score differential can be bounded resulting in an upper and lower bound. Thereafter, we discuss the boundary functions used to produce tight confidence sequences. Lastly, we present the confidence sequences and the corresponding null hypotheses. Proofs can be found in Choe & Ramdas (2022) and Howard et al. (2021). A more extensive derivation of confidence sequences is given in Choe & Ramdas (2022).

3.4.1 Cumulative Score Differential

Given two sequential forecasts a_i and b_i, $a_i, b_i \in [0,1]$, of a binary outcome y_i, we define the pointwise score differential $\delta_i := \mathbb{E}_{i-1}[S(a_i, y_i) - S(b_i, y_i)]$ to compare the quality of the forecasts where the expectation is taken over $y_i \sim \text{Bernoulli}(r_i)$. This parameter is however not observed because the choice of reality, r_i, is unknown. Thus, we define the empirical pointwise score differential as $\hat{\delta}_i := S(a_i, y_i) - S(b_i, y_i)$ which is observed by the statistician. The empirical pointwise score differential $\hat{\delta}_i$ is an unbiased estimator of δ_i.

To assess the quality of the forecasts over time, we define the average forecast score differential as

$$\Delta_t := \frac{1}{t}\sum_{i=1}^{T} \delta_i = \frac{1}{t}\sum_{i=1}^{T} \mathbb{E}_{i-1}[S(a_i, y_i) - S(b_i, y_i)], \tag{3}$$

and its empirical counterpart as

$$\widehat{\Delta}_t := \frac{1}{t}\sum_{i=1}^{T} \hat{\delta}_i = \frac{1}{t}\sum_{i=1}^{T} S(a_i, y_i) - S(b_i, y_i). \tag{4}$$

The objective is to measure how far $\widehat{\Delta}_t$ is from Δ_t while accounting for sampling uncertainty in y_t at each time t. We do this by introducing the cumulative score differential [1]

$$C_t := t(\widehat{\Delta}_t - \Delta_t) = \sum_{i=1}^{t} (\hat{\delta}_i - \delta_i), \quad i \geq 1, \tag{5}$$

[1] This is a cumulative differential of the score differential. For simplicity, we refer to this as the cumulative score differential and follow Choe & Ramdas (2022) in this regard.

where C_0 is set to one, $C_0 = 1$. This forms a martingale because

$$\begin{aligned}
\mathbb{E}_{t-1}[C_t] &= \mathbb{E}_{t-1}[\sum_{i=1}^{t}(\hat{\delta}_i - \delta_i)] \\
&= \mathbb{E}_{t-1}[(\hat{\delta}_t - \delta_t) + \sum_{i=1}^{t-1}(\hat{\delta}_i - \delta_i)] \\
&= \mathbb{E}_{t-1}[(\hat{\delta}_t - \delta_t)] + C_{t-1} \\
&= C_{t-1},
\end{aligned}$$

and because $\mathbb{E}_{t-1}[\hat{\delta}_t] = \delta_t$. Hereafter, we explain how the *sum process* $(C_t)_{t=0}^{\infty}$ can be uniformly bounded by exponential test supermartingales. By bounding C_t, we also bound the difference between $\widehat{\Delta}_t$ and Δ_t due to how C_t is defined in Equation 5. This allows us to construct Confidence Sequences.

3.4.2 Bounding Sum Process

To bound the *sum process* we introduce $\widehat{V}_t$ which measures the deviations of S_t from zero. This is also known as intrinsic time (Howard et al., 2020). Suppose that $|\hat{\delta}_i| \leq \frac{c}{2}, \forall i \geq 1$ for some $c > 0$. We know that the empirical pointwise score differential, $\hat{\delta}$, is bounded because we use the Brier score, $|\hat{\delta}| \leq 1$. Then define the variance process to be

$$\widehat{V}_t = \sum_{i=1}^{t}(\widehat{\delta}_i - \gamma_i)^2,$$

where $(\gamma)_{i=1}^{\infty}$ is any predictable sequence that lies in $[-\frac{c}{2}, \frac{c}{2}]$. Following Choe & Ramdas (2022), we choose γ_i to be the previous average score differential $\widehat{\Delta}_{i-1}$ such that the variance process takes the form of

$$\widehat{V}_t = \sum_{i=1}^{t}(\widehat{\delta}_i - \widehat{\Delta}_{i-1})^2.$$

We define a one-sided confidence bound $u_\alpha(\widehat{V}_t)$, which we explain later in more detail, as any function of $\widehat{V}_t$ that C_t exceeds with probability α as

$$P\left(\forall t \geq 1 : C_t \leq u_\alpha(\widehat{V}_t)\right) \geq 1 - \alpha, \tag{6}$$

which states that the sums are bounded from above by $u_\alpha(\widehat{V}_t)$ at all times with probability of at least $1 - \alpha$. A lower bound can be formed similarly for $(-C_t, \widehat{V}_t)_{t=0}^{\infty}$.

This uniform boundary based on the definitions of $(C_t, \widehat{V}_t)_{t=0}^{\infty}$ exist according to Howard et al. (2020, 2021) if for all $\lambda \in [0, \lambda_{max})$, the *exponential process* defined as $L_0 = 1$ and

$$L_t(\lambda) = \exp\left(\lambda C_t - \psi(\lambda)\widehat{V}_t\right),\ t \geq 1, \tag{7}$$

is a test supermartingale. Defining $C_t = \sum_{i=1}^{t}(\hat{\delta}_i - \delta_i)$ and $\widehat{V}_t = \sum_{i=1}^{t}(\hat{\delta}_i - \widehat{\Delta}_{i-1})^2$. For any $\lambda \in [0, \lambda_{max})$, the exponential process $L_t(\lambda)$ is a test supermartingale as

$$\begin{aligned}
\mathbb{E}_{t-1}[L_t(\lambda)] &= \mathbb{E}_{t-1}\left[\exp\left(\lambda C_t - \psi(\lambda)\widehat{V}_t\right)\right] \\
&= \mathbb{E}_{t-1}\left[\exp\left(\lambda \sum_{i=1}^{t}(\hat{\delta}_i - \delta_i) - \psi(\lambda)\sum_{i=1}^{t}(\hat{\delta}_i - \widehat{\Delta}_{i-1})^2\right)\right] \\
&= \mathbb{E}_{t-1}\left[\frac{\exp\left(\lambda \sum_{i=1}^{t-1}(\hat{\delta}_i - \delta_i)\right) \cdot \exp\left(\lambda(\hat{\delta}_t - \delta_t)\right)}{\exp\left(\psi(\lambda)\sum_{i=1}^{t-1}(\hat{\delta}_i - \widehat{\Delta}_{i-1})^2\right) \cdot \exp\left(\psi(\lambda)(\hat{\delta}_t - \widehat{\Delta}_{t-1})^2\right)}\right] \\
&= \mathbb{E}_{t-1}\left[\exp\left(\lambda \sum_{i=1}^{t-1}(\hat{\delta}_i - \delta_i) - \psi(\lambda)\sum_{i=1}^{t-1}(\hat{\delta}_i - \widehat{\Delta}_{i-1})^2\right) \cdot \frac{\exp\left(\lambda(\hat{\delta}_t - \delta_t)\right)}{\exp\left(\psi(\lambda)(\hat{\delta}_t - \widehat{\Delta}_{t-1})^2\right)}\right] \\
&= \mathbb{E}_{t-1}\left[L_{t-1}(\lambda) \cdot \frac{\exp\left(\lambda(\hat{\delta}_t - \delta_t)\right)}{\exp\left(\psi(\lambda)(\hat{\delta}_t - \widehat{\Delta}_{t-1})^2\right)}\right] \\
&= L_{t-1}(\lambda) \cdot \mathbb{E}_{t-1}\left[\frac{\exp\left(\lambda(\hat{\delta}_t - \delta_t)\right)}{\exp\left(\psi(\lambda)(\hat{\delta}_t - \widehat{\Delta}_{t-1})^2\right)}\right] \\
&\leq L_{t-1}(\lambda),
\end{aligned}$$

where $\psi : [0, \lambda_{max}) \to \mathbb{R}$ resembles a cumulant-generating function and controls the speed with which C_t can grow relative to $\widehat{V}_t$ (Howard et al., 2020). Choe & Ramdas (2022) have proven this results to hold as they showed that

$$\mathbb{E}_{t-1}\left[\frac{\exp\left(\lambda(\hat{\delta}_t - \delta_t)\right)}{\exp\left(\psi(\lambda)(\hat{\delta}_t - \widehat{\Delta}_{t-1})^2\right)}\right] \leq 1,$$

for $\psi(\lambda) = c^{-2}(-log(1-c\lambda)-c\lambda)$ with $\lambda \in [0, 1/c)$, which is the cumulant-generating function a rescaled centered exponential distribution with a scale parameter, $c > 0$. Therefore, there exists a uniform boundary for $(C_t, \widehat{V}_t)_{t=0}^{\infty}$.

3.4.3 Uniform Boundary

The choice for the boundary function $u(\cdot)$ affects how compact the CS is. Howard et al. (2021) state that the simplest boundary functions are linear functions of $\widehat{V}_t$. However,

compact confidence sequences are obtained with curved functions. Following Choe & Ramdas (2022) we use the conjugate-mixture (CM) boundary function derived by Howard et al. (2021) which is a curved boundary function. The CM boundary is defined as follows

$$u_\alpha^{CM}(v) := \sup\left\{s \in \mathbb{R} : m(s,v) < \frac{1}{\alpha}\right\}, \quad v \geq 0 \tag{8}$$

$$m(s,v) := \int \exp\{\lambda s - \psi(\lambda)v\} dF(\lambda), \tag{9}$$

which is crossed with probability $\alpha \in (0,1)$. The rationale behind this is as follows.

For any distribution F defined on $[0, \lambda_{max})$, the mixture $L_t^{mix} := \int L_t(\lambda) dF(\lambda)$ is a test supermartingale since $L_t(\lambda)$ is a test supermartingale for $\lambda \in [0, \lambda_{max})$. This result is proven by Howard et al. (2021). Therefore, $m(S_t, \widehat{V}_t) = L_t^{mix}$ is a test supermartingale. Using Ville's inequality, it holds that $P(\exists t \geq 1 : m(S_t, \widehat{V}_t) < 1/\alpha) \geq 1 - \alpha$. This in turn implies that $P(\exists t \geq 1 : S_t < u_\alpha^{CM}(v)) \geq 1 - \alpha)$. An appropriate choice for F yields a closed-form expression for L_t^{mix} similar to how an appropriately chosen prior distribution yields a closed-form expression for a posterior distribution. Choe & Ramdas (2022) report that the exponential ψ in combination with F as a gamma distribution yields a closed-from expression of u called the gamma-exponential mixture boundary. This gamma-exponential mixture boundary is given in Appendix A.3 of Howard et al. (2021).

3.4.4 Empirical Bernstein Confidence Sequences and Uniform Boundary

We will present here the Empirical Bernstein CS that we will use to compare sequential forecasters. This corresponds with Theorem 2 of Choe & Ramdas (2022) who have stated this for a more general setting where $\widehat{V}_t = \sum_{i=1}^t (\widehat{\delta}_i - \gamma_i)^2$, where $(\gamma)_{i=1}^\infty$ is any predictable sequence that lies in $[-\frac{c}{2}, \frac{c}{2}]$.

Let $u = u_\alpha^{CM}$ and $\widehat{V}_t = \sum_{i=1}^t (\widehat{\delta}_i - \widehat{\Delta}_{i-1})^2$. Then for any $\alpha \in (0,1)$

$$CS_t^{EB} := \left(\widehat{\Delta}_t \pm \frac{u(\widehat{V}_t)}{t}\right), \tag{10}$$

forms a $(1-\alpha)$ CS for Δ_t by Theorem 2 of Choe & Ramdas (2022).

Equation 10 is equivalent to stating that Δ_t is contained in the confidence sequence C_t^{EB} at all times with at least $1-\alpha$ probability. Choe & Ramdas (2022) define the corresponding null hypothesis as

$$\mathcal{H}_0(a,b) : \Delta_t \leq 0, \quad \forall t = 1, 2, ..., \tag{11}$$

which states that forecaster A is on average not any better than forecaster B. The null hypothesis that forecaster B is on average not any better than forecaster A is given as

$$\mathcal{H}_0(b, a) : \Delta_t \geq 0. \tag{12}$$

The null $\mathcal{H}_0(a, b)$ is rejected if the lower bound is positive, $\widehat{\Delta}_t - u(\widehat{V}_t)/t > 0$, while we reject $\mathcal{H}_0(b, a)$ if the upper bound is negative, $\widehat{\Delta}_t + u(\widehat{V}_t)/t < 0$.

4 Simulation Study

We study the advantages and disadvantages of the anytime-valid confidence sequences (CS) relative to the Diebold-Mariano (DM) and Giacomini-White (GW) tests. We do this by examining the following aspects.

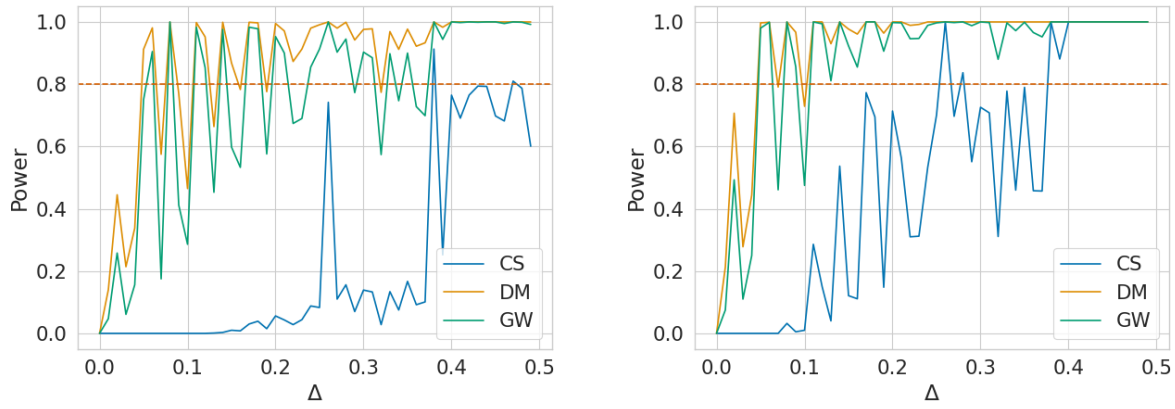
1. Power of CS, DM and GW tests

2. Anytime-validity of CS, DM and GW tests

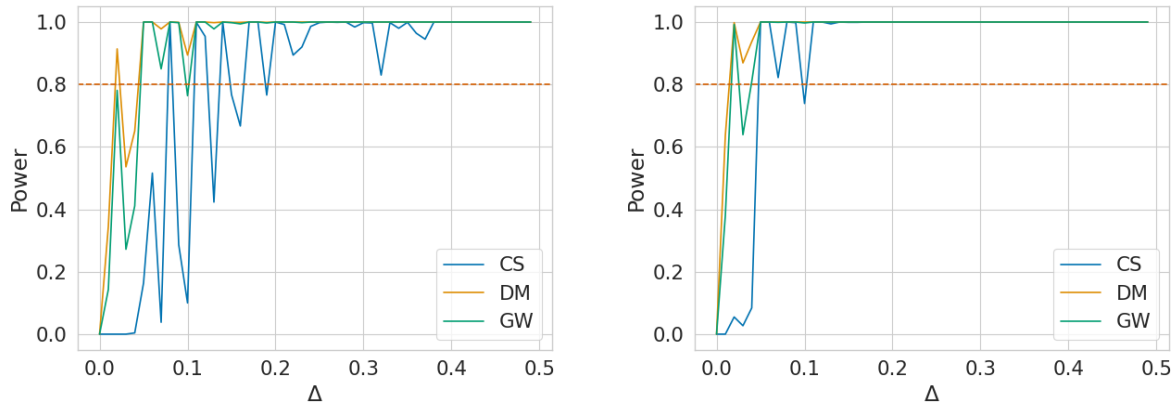
This will demonstrate to us where the benefits of anytime-valid methods lie relative to the DM and GW tests. In the power analysis, we measure the power of these tests in the case of a constant average score differential Δ. Afterwards, we introduce a trend in the score differential Δ and observe how much data these tests need to reject the null. This is motivated by the result of Choe & Ramdas (2022) which states that anytime-valid methods do not need larger sample sizes than the DM and GW test for high power. Lastly, we assess the impact of continuous monitoring on the parametric test and contrast that with CS. This shows the effect of continuous monitoring on the power and (cumulative) type I error of these tests. Here, we also contrast our results against the results of Choe & Ramdas (2022) and Henzi & Ziegel (2021) who conducted similar experiments.

4.1 Power Analysis

Choe & Ramdas (2022) report that anytime-valid methods do not require a larger sample size for high power. We will examine this through a power analysis. We simulate a Bernoulli outcome $Y_t \sim \text{Bernoulli}(p = 0.7)$ and select the forecasts a_t, b_t such that the true average score differential $\Delta_t = \frac{1}{t}\sum_{i=1}^{t} \mathbb{E}_{i-1}[S(a_i, y_i) - S(b_i, y_i)]$ takes values on the grid $\{0.01, 0.02, 0.03, ..., 0.48, 0.49, 0.50\}$. We generate 1000 samples and estimate the



(a) $T = 50$ (b) $T = 100$



(c) $T = 200$ (d) $T = 500$

Figure 2: Power of confidence sequences (CS), DM and GW test for a given score differential Δ for various sample sizes.

power. We examine the sample sizes of 100, 200 and 500 and contrast the power of CS with the power of the DM and GW tests.

Figure 2 displays the results. We find that the DM test has the highest power, followed by the GW test. Confidence sequences have the lowest power, which is expected. The power increases in the sample size and the score differential Δ. However, the probability of rejecting the null is 'volatile'. This volatility decreases in the sample size T. In general, CS needs more than twice as many observations as the DM test to reject the null for a given value of Δ. For example, the DM test needs 200 observations to reject the null with high power for $\Delta = 0.05$, while CS needs close to 500 observations.

This result is also seen in the following experiment, where we simulate data from a moving-average process. It displays the power of CS, and DM and GW tests and

resembles the experiment in Section 4.2 of Henzi & Ziegel (2021). We simulate data from the following moving-average process

$$Z_t = \epsilon_t + \theta \sum_{j=1}^{4} \epsilon_{t-j} \tag{13}$$

$$Y_t = \mathbb{1}\{Z_t > 0\}, \tag{14}$$

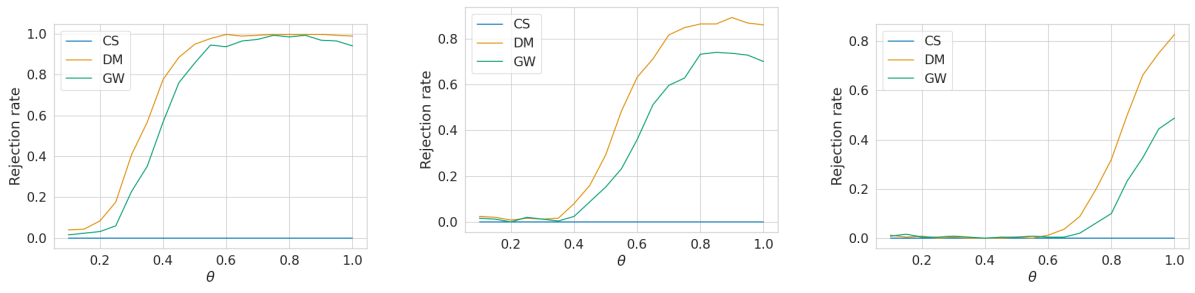
with ϵ following a standard normal distribution. We compare the following forecasts

$$a_{t,h} = P(Z_t > 0 | Z_{t-h}) \tag{15}$$

$$b_{t,h} = P(Z_t > 0 | Z_{t-h-1}), \tag{16}$$

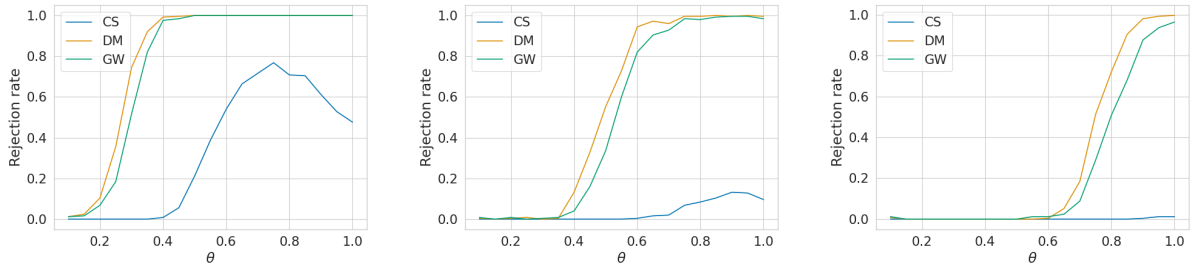
for lags $h = 1, 2, 3$. Given positive θ, $a_{t,h}$ outperforms $b_{t,h}$. The forecasting skill of $a_{t,h}$ and $b_{t,h}$ diverge for increasing θ and are equal when θ is zero. This data-generating process introduces an autocorrelation structure in the data. We generate 1000 samples and contrast the power of CS with the power of DM and GW tests.

Figures 3, 4 and 5 give the power curves for sample sizes 200, 500, and 600, respectively. Power decreases overall in lag h and increases in sample size T. The DM test has the highest power, closely followed by the GW test. Confidence sequences have considerably less power. For $\theta = 0.6$ and lag $h = 1$, CS need 600 observations to reject the null that q is on average not better than the forecaster p with high power, while the DM and GW tests need less than 200 observations. The same can be seen in Figures 3b and 5b. For lag $h = 2$ and $\theta = 0.9$, the DM test rejects the null with high power with only 200 observations, while CS do not achieve a power of 40% with 600 observations.



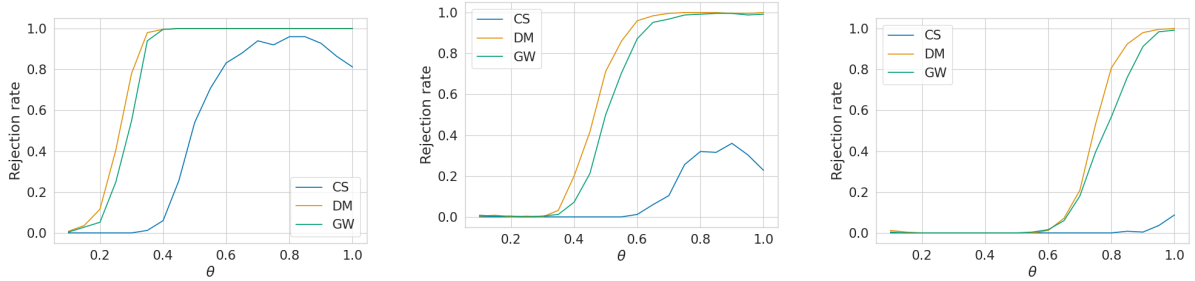
(a) $h = 1$ (b) $h = 2$ (c) $h = 3$

Figure 3: Power of confidence sequences (CS), DM and GW test with different lags h and $T = 200$



(a) $h = 1$ (b) $h = 2$ (c) $h = 3$

Figure 4: Power of confidence sequences (CS), DM and GW test with different lags h and $T = 500$



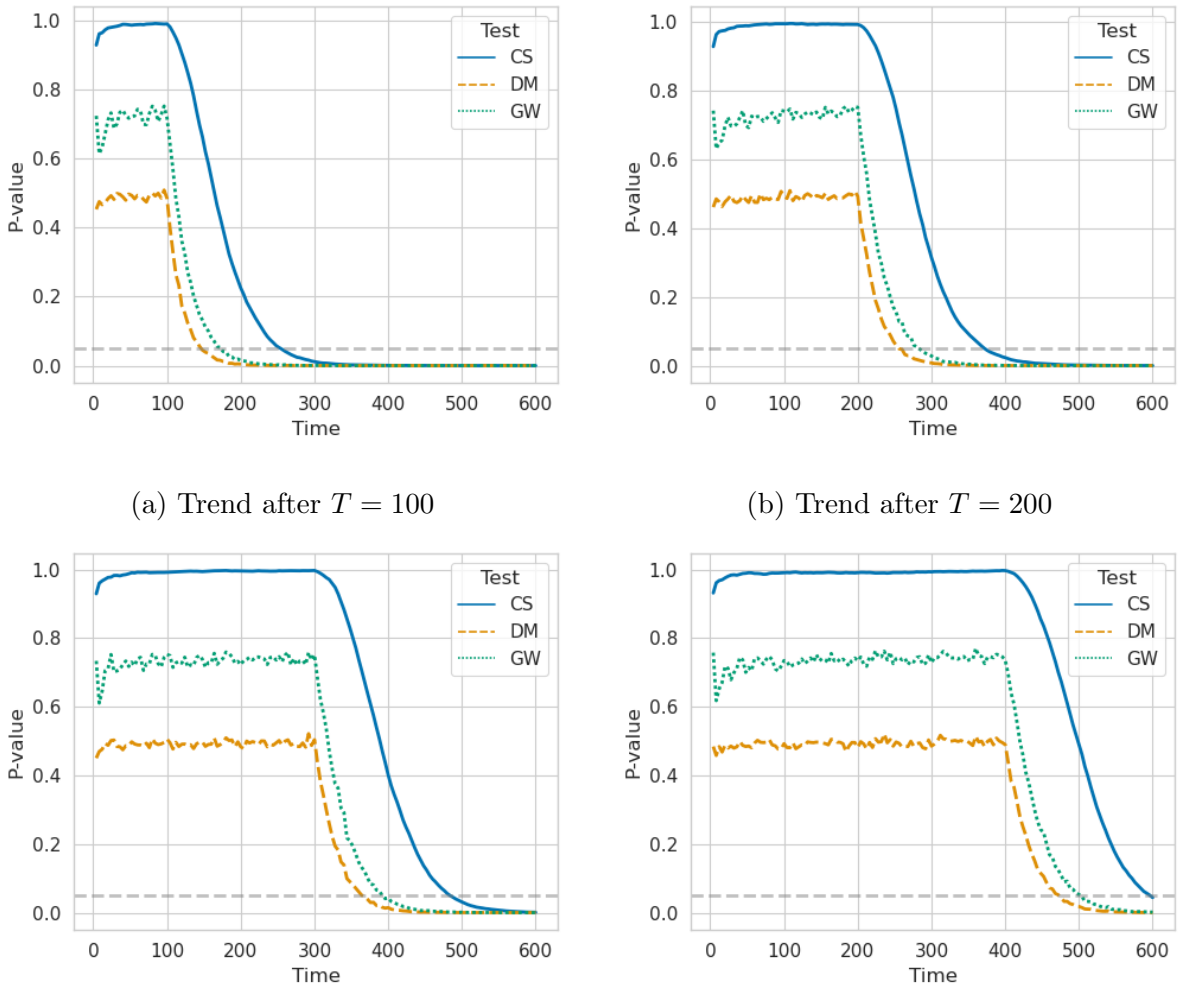
(a) $h = 1$ (b) $h = 2$ (c) $h = 3$

Figure 5: Power of confidence sequences (CS), DM and GW test with different lags h and $T = 600$

4.2 Detecting Trends in Forecast Performance

Choe & Ramdas (2022) concluded that CS does not need more data than DM and GW tests for high power from an experiment in which they introduced a trend in the score differentials. This setup violates the assumption that underlies the DM test. For the sake of completeness, the DM test and the corresponding null hypothesis can be found in Section A. The underlying assumption of the DM test is that the loss-differentials are assumed to be stationary. This means that the mean, variance and autocovariances are constant over time. We will perform a similar simulation as Choe & Ramdas (2022) and violate this assumption by introducing a trend in δ_t and compare how many observations the DM, GW and CS need to reject the null hypothesis that $\Delta_t \leq 0$. We simulate 1000 samples and compute the average p-value for each sample size $t \leq T$.

Choe & Ramdas (2022) used a sample size of 10.000 in their experiment. There, they simulated forecasts with a score differential equal to 0 for about $t \leq 7000$, afterwards a



(a) Trend after $T = 100$

(b) Trend after $T = 200$

(c) Trend after $T = 300$

(d) Trend after $T = 400$

Figure 6: P-values of the null that forecaster A is no better than forecaster B with a trend introduced at varying time points

trend was introduced in the score differential Δ. Such a setup makes sense if one has ample data but not for macroeconomic data as ten thousand observations correspond to 833 and 2500 years of monthly and quarterly observations, respectively. This resulted in very tight CS before the trend was introduced. In contrast, we will introduce a trend in Δ_t after 100, 200, 300 and 400 observations and only limit ourselves to a maximum sample size of 600 observations. This experiment will be more meaningful and generalizable to macroeconomic data.

Figure 6 displays the results. We find that the DM test is the first to reject the null hypothesis, followed by GW and CS. The number of observations that are needed for the DM and GW tests to reject the null is small. However, CS needs much more data to

reject the null. Table 1 displays the number of observations the DM, GW test and CS need to reject the null after the trend is introduced. We find that CS need approximately three times as much data as the DM test to reject the null and that the difference is more than 100 observations. In a setting with less than 600 observations, which frequently is the case when working with macroeconomic data, this is a significant difference. This corresponds to more than 25 years of quarterly data and roughly 8 years of monthly data.

Table 1: Number of observations needed to reject the null

	$T = 100$	$T = 200$	$T = 300$	$T = 400$
DM	48	60	64	72
GW	72	84	96	100
CS	160	176	192	200

4.3 Anytime-Validity

We assess the impact of continuous monitoring on the DM test and CS via two experiments.

In the first experiment, we assess the impact of continuous monitoring on the type I error and the power of the DM test, and contrast this with CS. We redo the first simulation of Henzi & Ziegel (2021). We simulate two forecasts $a_t, b_t \sim \text{Unif}(0, 1)$. Define a mixing-weight $\mu \in [0, 1]$. Then we define $\pi_t = \mu b_t + (1 - \mu)a_t$ and generate Bernoulli outcome Y_{t+1} with mean π_t. In this case, a_t is on average at least as good as b_t if and only if $\mu \leq 0.5$. The null hypothesis that we will be testing is

$$H_0 : \mathbb{E}[\Delta_t] \leq 0, \tag{17}$$

which holds for $\mu \leq 0.5$. The rejection rates for $\mu \leq 0.5$ constitute type I error whereas the rejection rates for $\mu > 0.5$ constitute the power of the tests. We consider the values in the grid {0, 0.05, 0.10, .., 0.95, 1.00} for μ the following sample sizes, {100, 200, 500}. Furthermore, we compare the rejection rates of CS with the rejection rates of the GW test and DM test with and without continuous monitoring.

We assess the impact of continuous monitoring on the type I error by examining the rejection rates for $\mu \leq 0.5$. In particular, we contrast the rejection rates of the DM test

with and without continuous monitoring by testing at $k = 1, 3, 5$ equally spaced points between $t = 1$ and $t = T$.

This simulation resembles the simulation in section 4.1 of Henzi & Ziegel (2021), but differs in the following aspect. In our simulation, we mainly contrast the DM test against confidence sequences. In contrast, Henzi & Ziegel (2021) contrast the t-test against e-values which correspond to the null hypothesis that forecaster A is at least as good as forecaster B at all times. This differs from the null hypothesis of confidence sequences, namely that forecaster B is an average not any better than forecaster A. A more detailed discussion about the differences between confidence sequences and e-values of Henzi & Ziegel (2021) can be found in Choe & Ramdas (2022).

Figure 7 displays the rejection rates of the CS, GW and DM tests with and without continuous monitoring, where the rejection rates for $\mu \leq 0.5$ constitute type I error and the rejection rates for $\mu > 0.5$ represent the power of these tests. We also present the rejection rates of the student's t-test to compare our findings with Henzi & Ziegel (2021).

In figure 7a, it is apparent that all tests have a type I error below the significance level of 5%, even at the boundary of $\mu = 0.5$. However, this is not the case for the DM test under continuous monitoring as shown in Figure 7b. We find that the type I error reaches 15% in the case of five stopping times between $t = 1$ and $t = T$. Even when tested only once between $t = 1$ and $t = T$, the DM test reports a higher type I error than the significance level, namely 8%. In contrast, the GW test does perform slightly better in this regard. We see from Figure 7c that at the boundary of $\mu = 0.5$, the rejection rate for five stopping times is slightly below 10%. Testing once between $t = 1$ and $t = T$ still yields a type I error below the 5% significance level. This implies that the continuous monitoring has a lesser impact on the type I error of the GW test.

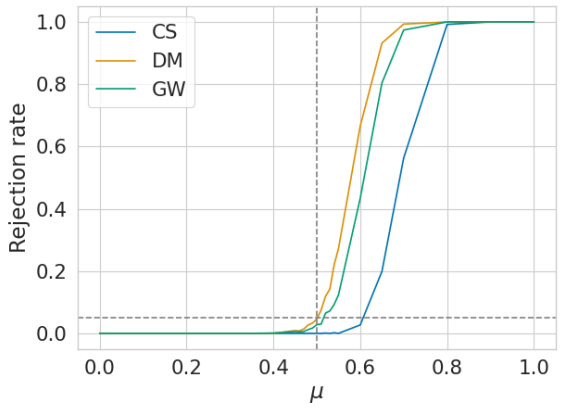
The student's t-test under continuous monitoring presented in Figure 7d yields similar rejection rates as the DM test. The type I error also reaches 15% in the case of continuous monitoring with 5 stopping times, which resembles the findings of Henzi & Ziegel (2021). In fact, the DM test and the t-test yield the same rejection rates as can be seen in Figure 7e because the score differential Δ is not autocorrelated. The DM test adds a correction for autocorrelation to the variance, see Equation 22 in Section A. In the case of no autocorrelation, the DM test equals a t-test testing if the the score differentials have zero mean. Moreover, we find that the DM test has the most power followed by the

GW test. Confidence sequences have the least power. Furthermore, the DM test with continuous monitoring has more power (and a higher type I error). The same holds for the t-test and GW tests with continuous monitoring.

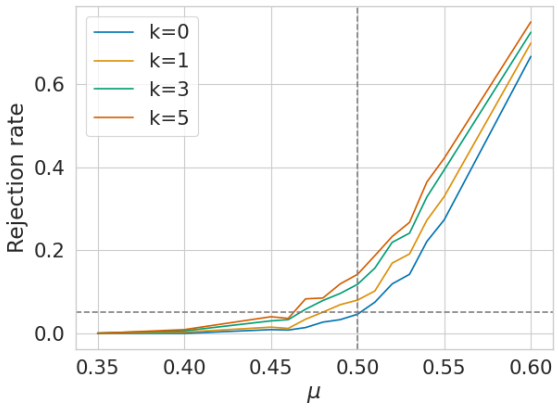
The second experiment contrasts the evolution of the cumulative Type I error of the DM and GW tests with that of confidence sequences over time. We simulate a Bernoulli outcome Y_t and simulate the forecasts a_t, b_t such that the true average score differential $\Delta_t = 0$. We estimate the cumulative type I error given as $P(\exists i \leq t : a_i \leq \alpha)$. According to Choe & Ramdas (2022), for CS this is equivalent to the cumulative miscoverage rate $P(\exists i \leq t : \Delta_t \notin C_i)$, where C_t is the confidence at time t.

This simulation is similar to the simulation of Choe & Ramdas (2022), but differs in the following. We calculate the cumulative type I error using a smaller sample than Choe & Ramdas (2022) who used a sample size of 10,000 observations. This corresponds to 833 and 2500 years of monthly and quarterly observations, respectively. In contrast, we use a sample size of 600 observations, which corresponds to 50 and 150 years of monthly and quarterly observations, respectively. The first 50 observations are used as a burn-in period such that the DM test statistic converges to its asymptotic distribution. Choe & Ramdas (2022) use a burn-in of 100 observations.

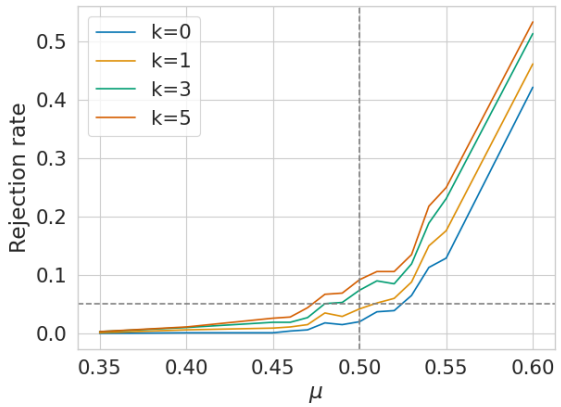
Figure 8 displays the results. In this experiment, $\Delta_t = 0$ for $t \geq 1$, indicating that the null hypothesis, $H_0 : \Delta_t \leq 0$, holds. We find that the cumulative type I error surpasses the significance level $\alpha = 0.05$ after 4 observations and reach a cumulative type I error of 27%. The GW test surpasses the significance level after 16 observations and reports a lower cumulative type I error than the DM test, namely 20%. In contrast, confidence sequences report a cumulative type I error of 0% for all observations. Choe & Ramdas (2022) attribute this to the fact that CSs are constructed using supermartingales and not martingales. Additionally, Choe & Ramdas (2022) report in their experiment that the DM test exceeds the significance level of 5% after 100 observations. However, in their setup, they test at intervals of 100 observations. We report that the DM test exceeds the significance level after 4 observations because we test at intervals of four observations. This corresponds to testing every quarter in case of monthly data or every year in case of quarterly data. Moreover, our results differ in the following aspect. We report that the DM test achieves a higher cumulative type I error than the GW test, while Choe & Ramdas (2022) reports the opposite.



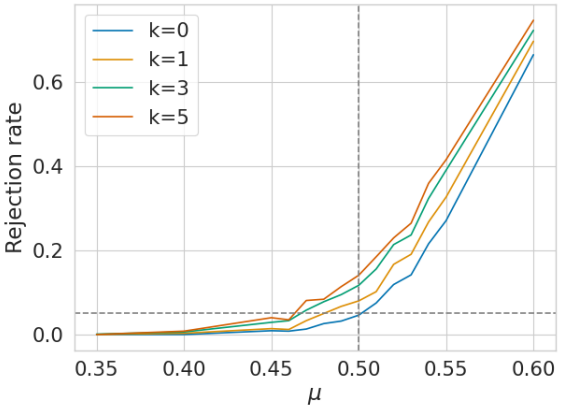
(a) Rejection rates of CS, DM and GW tests



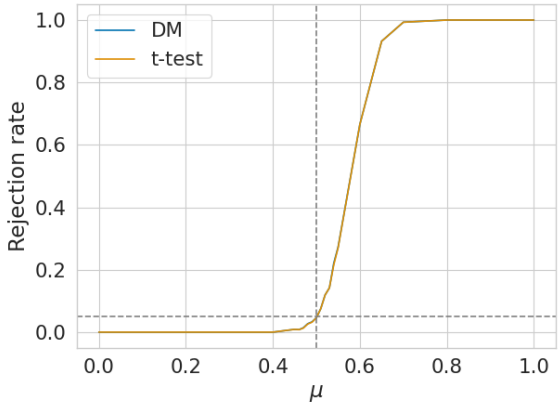
(b) Rejection rate DM with optional stopping



(c) Rejection rates GW with optional stopping



(d) Rejection rates t-test with optional stopping



(e) Rejection rate of DM and t-test

Figure 7: Rejection rates of CS, DM, GW and t-test for different values of μ with a sample size of $T = 600$. For $\mu \leq 0.5$ the rejection rates represent the type I error as the null hypothesis that p_t outperforms q_t holds. For $\mu > 0.5$, the alternative hypothesis holds and the rejection rates represent the power of the tests. Similar to Henzi & Ziegel (2021), we test at k equispaced points between $t = 1$ and $t = T$.

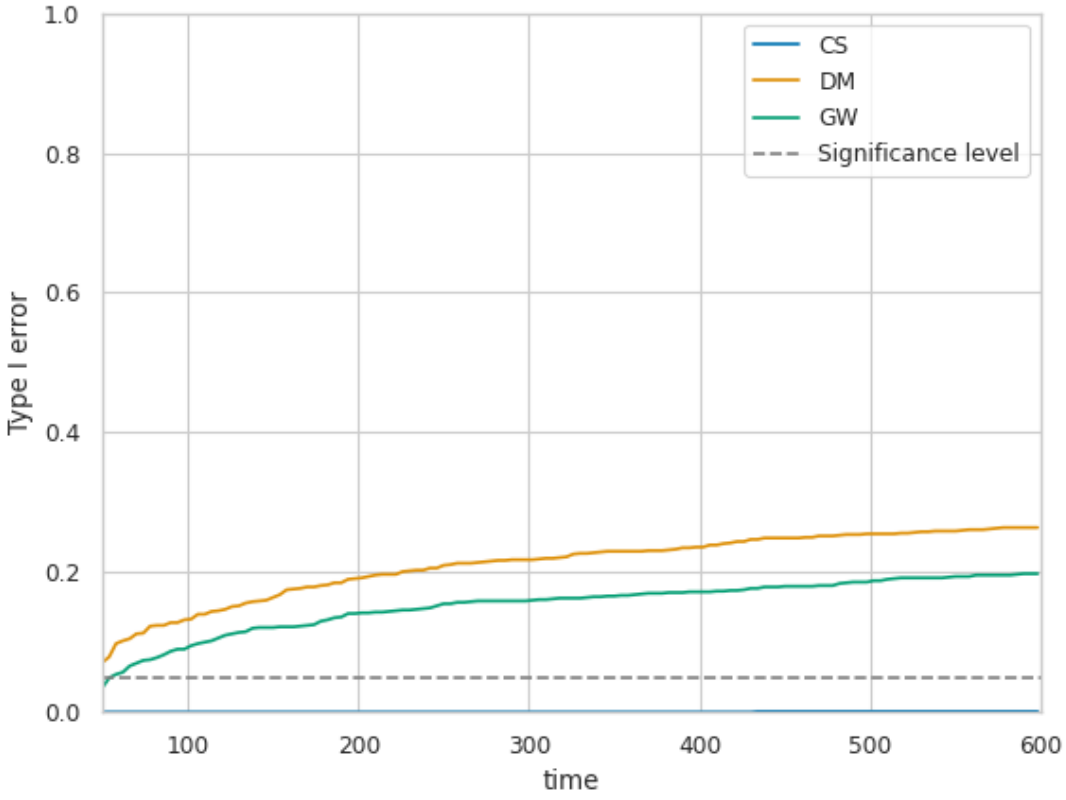


Figure 8: Cumulative type I error of CS, DM and GW tests

5 Empirical Application

In this section, we utilise anytime-valid confidence sequences to compare directional forecasts of US GDP growth and inflation. The forecasts are obtained from consumer sentiment, professional forecasters' and economists' expectations. We first describe our data set and afterwards, we present our results.

5.1 Data

We compare the directional forecasts of US GDP growth and inflation using anytime-valid confidence sequences. We analyse three directional forecasts constructed from consumer sentiment, professional forecasters and economists' forecasts. These are obtained from The Surveys of Consumers conducted by the Survey Research Center at the University of Michigan, the Survey of Professional Forecasters published by the Federal Reserve Bank of Philadelphia and the Livingston Survey respectively.

The Survey of Consumers is a monthly survey where a minimum of 600 respondents are asked about their attitudes and expectations concerning their personal finances, business conditions and buying conditions. This information is summarised in the Index of Consumer Sentiment (forecasts will be referred to as *sentiment* forecast). The survey has been conducted since 1946 and monthly and quarterly indices are published since 1978 and 1960 respectively, resulting in 545 monthly and 253 quarterly consumer sentiment indices. Figure 9 displays the evolution of the quarterly Index of Consumer Sentiment over time.

The Survey of Professional Forecasters (forecasts will be referred to as *professional* forecast) is a quarterly forecast conducted since 1968. In it, respondents are asked about their expectations and forecasts of macroeconomic variables, including unemployment, inflation, GDP, bond yields, and consumption expenditures. We use the mean growth forecasts concerning GDP and inflation. The forecast horizons[2] are the quarter in which the survey is conducted and the four subsequent quarters. This results in 219 forecasts. Figure 10 plots the 1-quarter ahead growth forecast against GDP growth over time.

The Livingston survey surveys economists's expectations twice a year since 1946. The respondents are asked about their forecasts of variables concerning, amongst others, GDP, prices and investment. The mean and median of the growth rate of these variables as well as the mean and median responses of the level are given. The respondents submit their forecasts for different horizons of which we use the period from two quarters beyond the survey date to four quarters beyond the survey date. These forecasts number in total 155. The relevant variables for this research are the mean and median growth rate forecast of GDP and CPI. Forecasts will be referred to as *economists* forecast.

Directional forecasts were constructed as follows. Strictly positive growth rate forecasts were assigned the value 1. Growth rate forecasts that were zero or negative were assigned the value 0 following Vrontos et al. (2021). Missing values were not imputed. In the case of the Survey of Professional Forecasters and the Livingston Survey, the mean growth rates made available were used. To facilitate the comparison between forecasts constructed from the Consumer Sentiment Index or Survey of Professional Forecasters with those constructed from the Livingston Survey, the quarterly growth rates are con-

[2]For the inflation rate, the forecast horizons given are: the previous, current and subsequent four quarters as well as the annual inflation rate for the current and subsequent year. However, we only limit ourselves to the quarterly forecasts of the current and subsequent four quarters.

Table 2: Summary statistics directional forecasts

	growth forecast		inflation forecast	
	mean	Brier score	mean	Brier score
Quarterly *professional*	0.98	0.97	0.92	0.84
Biannual *professional*	0.98	0.99	0.96	0.99
economist (biannual)	1.00	0.97	1.00	0.98
Biannual *sentiment*	0.52	0.49	0.48	0.50
Quarterly *sentiment*	0.47	0.49	0.50	0.59

solidated into biannual growth rates. This is explained in more detail in Section B of the Appendix.

Lastly, we present some statistics about these directional forecasts in Table 2. We notice that the means are very close or equal to 1 for the growth and inflation forecasts constructed from the Survey of Professional Forecasters and the Livingston Survey. These forecasts also achieve high Brier scores. This implies that these forecasts mostly predict positive outcomes and directional data mostly consist of positive outcomes. The consumer sentiment forecasts have a considerably lower mean and Brier score.

5.2 Consumers vs. Professional Forecasters

In our first sequential comparison, we compare the directional forecasts based on the Consumer Sentiment Index (*sentiment* forecasts) against directional forecasts based on the Survey of Professional Forecasters (*professional* forecasts). The growth forecasts that we compare span 218 quarters from the fourth quarter of 1968 until the first quarter of 2023. The inflation forecasts start in the third quarter of 1981 and end in the first quarter of 2023 and number 167 forecasts in total.

Figure 11 displays the results of sequentially comparing the *sentiment* forecasts against the *professional* forecasts. We find that the *professional* growth and inflation forecasts on average outperform the *sentiment* growth and inflation forecasts. Regarding the growth forecasts, the *professional* forecasts constantly outperform the *sentiment* forecasts as shown in Figure 11a, where the lower bound of the confidence sequence does not decrease. While the *professional* inflation forecasts on average outperform the *sentiment*

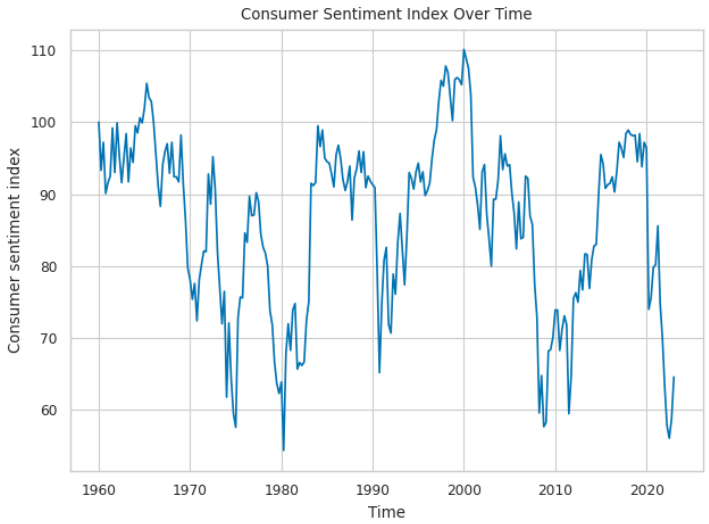


Figure 9: Consumer sentiment index over time

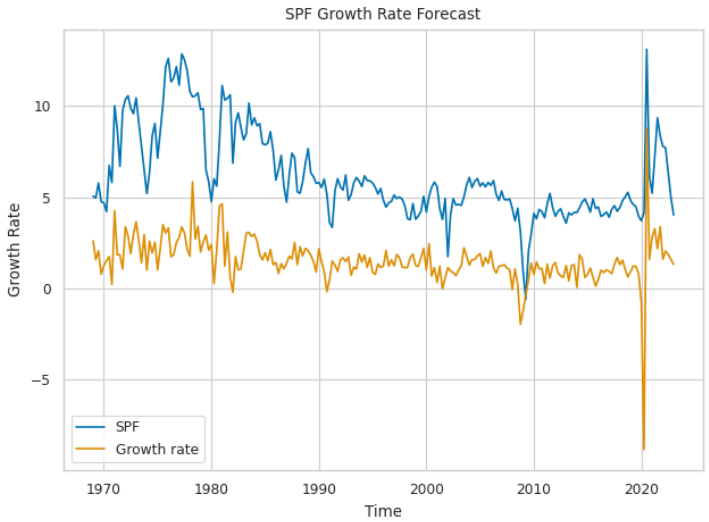
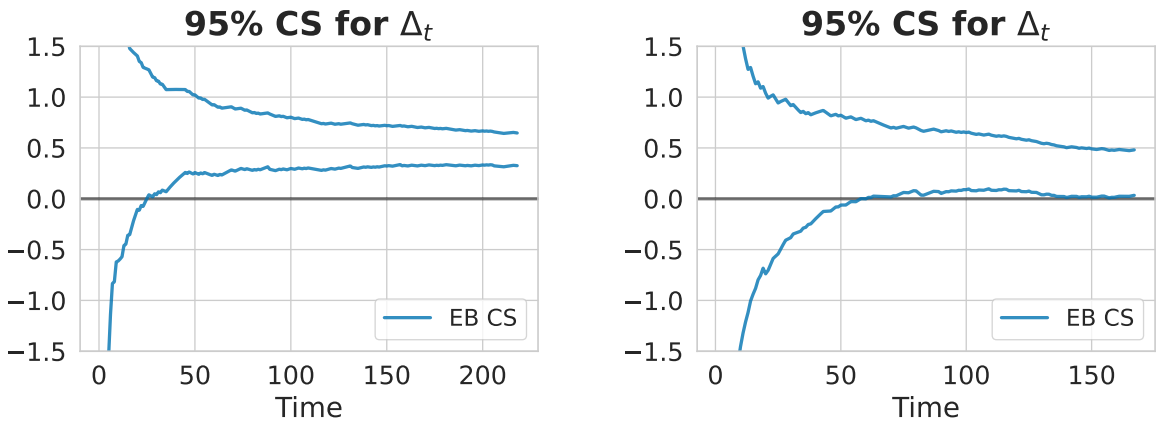


Figure 10: SPF growth rate forecast over time



(a) Growth forecasts

(b) Inflation forecasts

Figure 11: Consumer sentiment vs professional forecasters

Table 3: The confidence interval and Brier score of *sentiment* and *professional* forecasters evaluated at T = 167 for inflation forecasts and T = 217 for growth forecasts

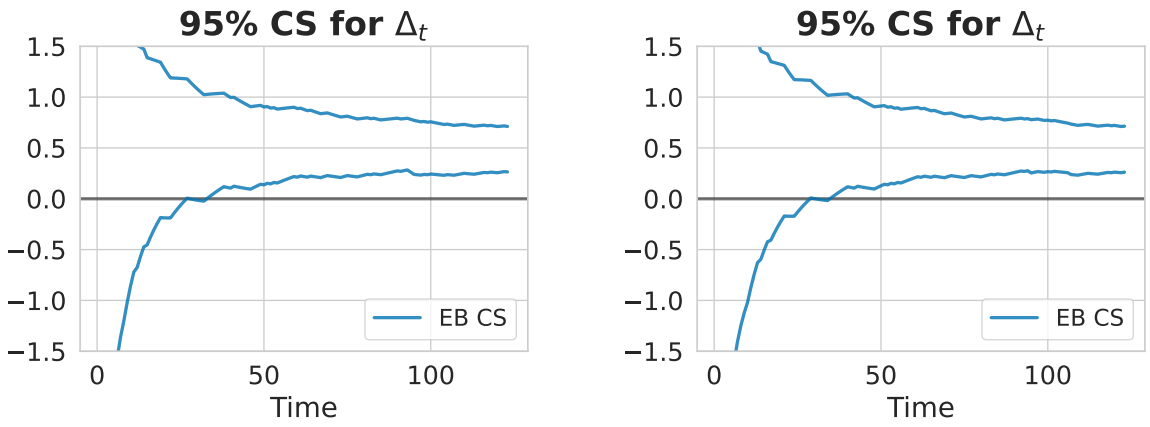
Variable	Confidence Interval	Brier score *sentiment*	Brier score *professional*
Growth	(0.325, 0.647)	0.49	0.97
Inflation	(0.034, 0.481)	0.59	0.84

inflation forecasts after 60 quarters, the performance of the *professional* inflation forecasts relative to *sentiment* forecasts drops around 120 quarters as is shown in Figure 11b. This can be seen in the decrease in the lower bound of the confidence sequence.

Furthermore, analysing the average Brier scores of the *sentiment* forecasts and *professional* forecasts given in Table 3, we find that the *sentiment* forecasts are more capable of forecasting the direction of inflation than the direction of GDP growth. This results in a higher Brier score. This finding is in line with economic theory that states that consumer expectations play a large role in inflation. The opposite is the case for the *professional* forecasts. These forecast the direction of growth better than the direction of inflation.

5.3 Consumers vs. Economists

Secondly, we compare the *sentiment* forecasts against the directional forecasts constructed from economists' expectations as published in the Livingston Survey (*economists* forecasts). To facilitate the comparison, the *sentiment* forecasts are aggregated into bi-



(a) Growth forecasts

(b) Inflation forecasts

Figure 12: Consumer sentiment vs economists' expectations

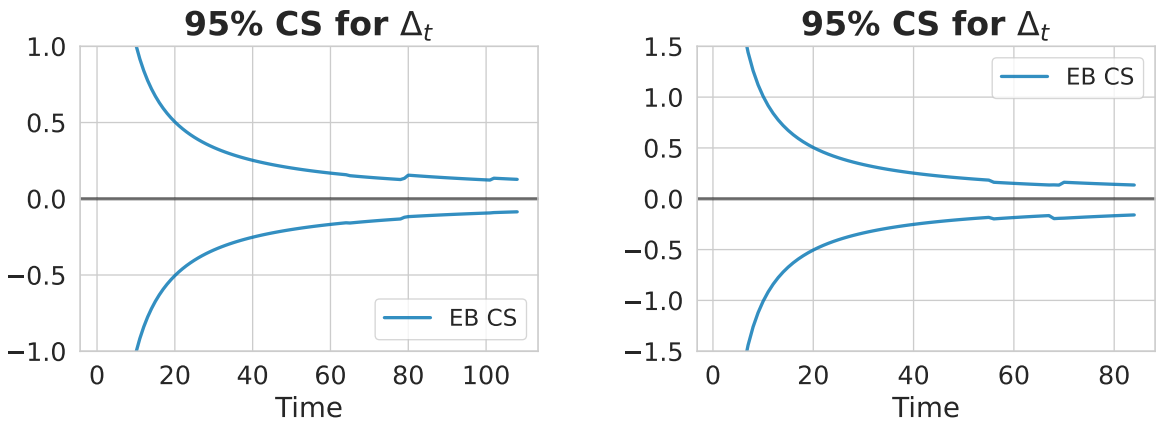
Table 4: The confidence interval and Brier score of *sentiment* and *economist* forecasters evaluated at T = 123 for inflation forecasts and T = 167 for growth forecasts

Variable	Confidence Interval	Brier score *sentiment*	Brier score *economist*
Growth	(0.264, 0.711)	0.49	0.98
Inflation	(0.262, 0.714)	0.5	0.98

annual forecasts. The growth and inflation forecasts span 60 years from 1962 until 2023 and number 123 forecasts in total.

The confidence sequences are given in Figure 12. Once more, the *sentiment* forecasts are outperformed. This time, both the growth and inflation forecasts of the *economists* forecasts consistently outperform the *sentiment* forecasts. This can be seen in the lower bound of the confidence sequence which is positive and does not decrease overall.

The Brier score in Table 4 illustrates the large gap in performance with the *economists* forecasts obtaining a high Brier score of 0.98. The *sentiment* forecasts achieve a much lower Brier score. In contrast with the quarterly *sentiment* forecasts, the biannual *sentiment* forecasts achieve a similar Brier score for the growth and inflation forecasts. The same holds for the *economists* forecasts.



(a) Growth forecasts

(b) Inflation forecasts

Figure 13: Professional forecasters vs economists's expectations

Table 5: The confidence interval and Brier score of *professional* and *economist* forecasters evaluated at T = 84 for inflation forecasts and T = 108 for growth forecasts

Variable	Confidence Interval	Brier score *professional*	Brier score *economist*
Growth	(-0.086, 0.127)	0.99	0.97
Inflation	(-0.159, 0.135)	0.99	0.98

5.4 Professional Forecasters vs. Economists

Next, we compare the *professional* forecasts against *economists* forecasts. In the two previous comparisons, we have seen that both had high Brier scores and therefore on average outperformed the *sentiment* forecasts. To facilitate the comparison, the *professional* forecasts are aggregated into biannual forecasts. The forecasts growth span 54 years from 1969 until 2023 and number 108 forecasts in total. The inflation forecasts start in 1981 and number 84 forecasts in total.

The results of the sequential comparison are given in Figure 13. We find that neither forecasts outperform the other on average. We also note that the widths of the confidence sequence in Figure 13b increase around 65 observations. This can occur if the variance of the cumulative score difference increases when the pointwise score differentials $\hat{\delta}_i$ differ vastly from the previous period's average score differential $\widehat{\Delta}_{i-1}$.

The average Brier score of the *professional* forecasts and *economists* forecasts is given

in Table 5. We notice that both forecasts achieve high Brier scores close to 1 and both forecasts achieve similar Brier scores for the growth and inflation forecasts. In contrast, the quarterly *professional* forecasts in Table 3 predict the direction of growth better than the direction of inflation.

6 Conclusion

We have compared the power and anytime-validity of confidence sequences, DM and GW tests to weigh the advantages and disadvantages of confidence sequences vis-a-vis DM and GW tests in small samples. We find that confidence sequences have less power: confidence sequences need more than twice as much data as the DM test to reach the same power and it needs more than three times as much data to detect a change in forecaster performance. On the other hand, confidence sequences are anytime-valid. In contrast, the DM and GW tests suffer from an inflated type I error in the case of continuous testing. Our empirical application in directional forecasts showed that confidence sequences can detect if the performance of forecasters changed in the past. In light of our findings, we conclude that Save Anytime-Valid Inference (SAVI) methods like confidence sequences lack the power to be useful in 'live' testing but can test 'backwards' to detect changes in past forecaster performance.

A promising direction of future research is to improve the tightness of confidence sequences by examining different choices for the sequence $(\gamma)_{i=1}^{\infty}$. As Choe & Ramdas (2022) stated: a smarter choice may lead to tighter CS. This in turn increases the power of confidence sequences.

As of now, SAVI methods are best suited for data-rich scenarios. Examples of these include exchange rate data, stock market data and any other data that is available at a daily or even hourly level. These contain enough data to offset the loss in power that comes with anytime-valid methods and thus can utilise the anytime-validity of SAVI methods.

References

Ashley, R., Granger, C. W. J. & Schmalensee, R. (1980). Advertising and aggregate consumption: An analysis of causality. *Econometrica*, *48*(5), 1149–1167. Retrieved 2023-05-26, from http://www.jstor.org/stable/1912176

Blaskowitz, O. & Herwartz, H. (2011). On economic evaluation of directional forecasts. *International Journal of Forecasting*, *27*(4), 1058-1065. Retrieved from https://www.sciencedirect.com/science/article/pii/S0169207011000057 doi: https://doi.org/10.1016/j.ijforecast.2010.07.002

Blaskowitz, O. & Herwartz, H. (2014). Testing the value of directional forecasts in the presence of serial correlation. *International Journal of Forecasting*, *30*(1), 30-42. Retrieved from https://www.sciencedirect.com/science/article/pii/S0169207013000873 doi: https://doi.org/10.1016/j.ijforecast.2013.06.001

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*(1), 1 – 3.

Choe, Y. J. & Ramdas, A. (2022). *Comparing sequential forecasters.*

Darling, D. A. & Robbins, H. (1967). Confidence sequences for mean, variance, and median. *Proceedings of the National Academy of Sciences*, *58*(1), 66–68.

Diebold, F. X. (2015). Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of diebold–mariano tests. *Journal of Business & Economic Statistics*, *33*(1), 1-1. Retrieved from https://doi.org/10.1080/07350015.2014.983236 doi: 10.1080/07350015.2014.983236

Diebold, F. X. & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, *13*(3), 253-263. Retrieved from https://

www.tandfonline.com/doi/abs/10.1080/07350015.1995.10524599 doi: 10.1080/07350015.1995.10524599

Fisher, R. A. (1922). On the interpretation of χ^2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, *85*(1), 87–94. Retrieved 2023-05-01, from http://www.jstor.org/stable/2340521

Franses, P., van Dijk, D. & Opschoor, A. (2014). *Time series models for business and economic forecasting, 2nd edition*. Cambridge University Press.

Giacomini, R. & White, H. (2006). Tests of conditional predictive ability. *Econometrica*, *74*(6), 1545–1578. Retrieved 2023-05-02, from http://www.jstor.org/stable/4123083

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

Good, I. (1971). Comment on "measuring information and uncertainty" by robert j. buehler. *Foundations of Statistical Inference*, 337–339.

Good, I. J. (1952). Rational Decisions. *Journal of the Royal Statistical Society. Series B (Methodological)*, *14*(1), 107–114. Retrieved 2023-03-28, from https://www.jstor.org/stable/2984087 (Publisher: [Royal Statistical Society, Wiley])

Granger, C. W. J. & Pesaran, M. H. (2000). Economic and statistical measures of forecast accuracy [Article]. *Journal of Forecasting*, *19*(7), 537 – 560. Retrieved from https://www.scopus.com/inward/record.uri?eid=2-s2.0-0000109477&doi=10.1002%2f1099-131x%28200012%2919%3a7%3c537%3a%3aaid-for769%3e3.3.co%3b2-7&partnerID=40&md5=f569e668e5888de3da8b8fe6f03959e2 (Cited by: 224; All Open Access, Green Open Access) doi: 10.1002/1099-131x(200012)19:7⟨537::aid-for769⟩3.3.co;2-7

Henzi, A. & Ziegel, J. F. (2021, 09). Valid sequential inference on probability forecast performance. *Biometrika*, *109*(3), 647-663. Retrieved from https://doi.org/10.1093/biomet/asab047 doi: 10.1093/biomet/asab047

Hewamalage, H., Ackermann, K. & Bergmeir, C. (2023). Forecast evaluation for data scientists: common pitfalls and best practices. *Data Mining and Knowledge Discovery*, *37*(2), 788–832. Retrieved 2023-03-29, from https://doi.org/10.1007/s10618-022-00894-5 doi: 10.1007/s10618-022-00894-5

Howard, S. R., Ramdas, A., McAuliffe, J. & Sekhon, J. (2020). Time-uniform Chernoff bounds via nonnegative supermartingales. *Probability Surveys*, *17*(none), 257 – 317. Retrieved from https://doi.org/10.1214/18-PS321 doi: 10.1214/18-PS321

Howard, S. R., Ramdas, A., McAuliffe, J. & Sekhon, J. (2021). Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, *49*(2), 1055 – 1080. Retrieved from https://doi.org/10.1214/20-AOS1991 doi: 10.1214/20-AOS1991

Jamieson, K. & Jain, L. (2018). A bandit approach to multiple testing with false discovery control. In *Proceedings of the 32nd international conference on neural information processing systems* (p. 3664–3674). Red Hook, NY, USA: Curran Associates Inc.

Jamieson, K., Malloy, M., Nowak, R. & Bubeck, S. (2014, 13–15 Jun). lil' ucb : An optimal exploration algorithm for multi-armed bandits. In M. F. Balcan, V. Feldman & C. Szepesvári (Eds.), *Proceedings of the 27th conference on learning theory* (Vol. 35, pp. 423–439). Barcelona, Spain: PMLR. Retrieved from https://proceedings.mlr.press/v35/jamieson14.html

Jennison, C. & Turnbull, B. W. (1984). Repeated confidence intervals for group sequential clinical trials. *Controlled Clinical Trials*, *5*(1), 33–45.

Jennison, C. & Turnbull, B. W. (1989). Interim analyses: The repeated confidence interval approach. *Journal of the Royal Statistical Society. Series B (Methodological)*, *51*(3), 305–361. Retrieved 2023-05-01, from http://www.jstor.org/stable/2345448

Johari, R., Pekelis, L. & Walsh, D. J. (2019). *Always valid inference: Bringing sequential analysis to a/b testing.*

Lai, T. L. (1976). Boundary crossing probabilities for sample sums and confidence sequences. *The Annals of Probability*, *4*(2), 299–312.

Lai, T. L., Gross, S. T. & Shen, D. B. (2011). Evaluating probability forecasts. *The Annals of Statistics*, *39*(5), 2356 – 2382. Retrieved from https://doi.org/10.1214/11-AOS902 doi: 10.1214/11-AOS902

Pesaran, H. & Timmermann, A. (1992, 10). A simple nonparametric test of predictive performance. *Journal of Business & Economic Statistics*, *10*, 561-65. doi: 10.1080/07350015.1992.10509922

Pesaran, M. H. & Timmermann, A. (2009). Testing dependence among serially correlated multicategory variables. *Journal of the American Statistical Association*, *104*(485), 325-337. Retrieved from https://doi.org/10.1198/jasa.2009.0113 doi: 10.1198/jasa.2009.0113

Robbins, H. (1970). Statistical methods related to the law of the iterated logarithm. *The Annals of Mathematical Statistics*, *41*(5), 1397–1409.

Shafer, G., Shen, A., Vereshchagin, N. & Vovk, V. (2011). Test martingales, bayes factors and p-values. *Statistical Science*, *26*(1), 84–101. Retrieved 2023-07-11, from http://www.jstor.org/stable/23059157

Stekler, H. O. (1991). Macroeconomic forecast evaluation techniques. *International Journal of Forecasting*, *7*(3), 375–384. Retrieved 2023-03-28, from https://www.sciencedirect.com/science/article/pii/016920709190011J doi: 10.1016/0169-2070(91)90011-J

Tsuchiya, Y. (2013). Are government and imf forecasts useful? an application of a new market-timing test. *Economics Letters*, *118*(1), 118–120.

Tsuchiya, Y. (2016). Do production managers predict turning points? a directional analysis. *Economic Modelling*, *58*, 1–8.

Ville, J. (1939). *Étude critique de la notion de collectif.* Retrieved from http://eudml.org/doc/192893

Vrontos, S. D., Galakis, J. & Vrontos, I. D. (2021). Implied volatility directional forecasting: a machine learning approach. *Quantitative Finance*, *21*(10), 1687-1706. Retrieved from https://doi.org/10.1080/14697688.2021.1905869 doi: 10.1080/14697688.2021.1905869

Wilcoxon, F. (1947). Probability tables for individual comparisons by ranking methods. *Biometrics*, *3*(3), 119–122. Retrieved 2023-05-26, from `http://www.jstor.org/stable/3001946`

A Diebold-Mariano Test Statistic

We use the following DM test statistic, which Franses et al. (2014) state is the most popular in practice.

The loss-differential d_t defined as

$$d_t := e^2_{A,t|t-h} - e^2_{B,t|t-h} \tag{18}$$

$$:= \hat{\delta}_t, \tag{19}$$

where $e^2_{A,t|t-h}$ and $e^2_{B,t|t-h}$ denote the forecast errors of forecasters A and B, respectively. The null hypothesis states that the forecast errors of forecaster A and B are equal. This implies $H_0 : \mathbb{E}[d_t] = 0$. The test statistic is given as

$$\frac{\bar{d}}{\sqrt{\hat{\sigma}^2_{d_t}/T}} \sim N(0,1), \tag{20}$$

where $\bar{d} = \frac{1}{T}\sum_{t=1}^{T} d_t$ denotes the sample mean of the loss-differential and $\hat{\sigma}^2_{d_t}$ is the variance of d_t. This variance is computed as

$$\hat{\sigma}^2_{d_t} = \hat{\gamma}_0 + 2\sum_{j=0}^{h-1} \hat{\gamma}_j, \tag{21}$$

where the forecasting horizon h equals 1 and the j-th order sample autocovariance, $\hat{\gamma}_j$, is computed as

$$\hat{\gamma}_j = \frac{1}{T}\sum_{t=0}^{T-1-j} (d_t - \bar{d})(d_{t-j} - \bar{d}). \tag{22}$$

B Aggregation of Quarterly Forecasts to Biannual Forecasts

The Index of Consumer Sentiment is transformed to biannual forecasts as follows. First, we compute the biannual Index of Consumer Sentiment by computing the mean of the

first two and last two quarters of the quarterly index, which correspond with the first and second half of a year, respectively. This is then transformed to directional variables. If the biannual index at time t is greater than the index at time $t-1$, then the directional variables is assigned one. Otherwise, the variable takes on zero.

The forecasts of the Survey of Professional Forecasters are growth forecasts. These are transformed into biannual forecasters in the following way. First, the percentages are transformed to decimals, $x/100 + 1$. Afterward, the growth forecasts adjacent quarters (such as the first and second quarters, as well as the third and fourth quarters) are combined using the following formula:

$$(g_q * g_{q-1} - 1)/100,$$

where g_q is the growth forecast in quarter q, with $q = 2, 4$. These biannual growth rates are then transformed to directional variables. In this transformation, positive growth rates are assigned the value 1, while negative growth rates are assigned the value 0.

C Power Analysis - Bernoulli(0.4)

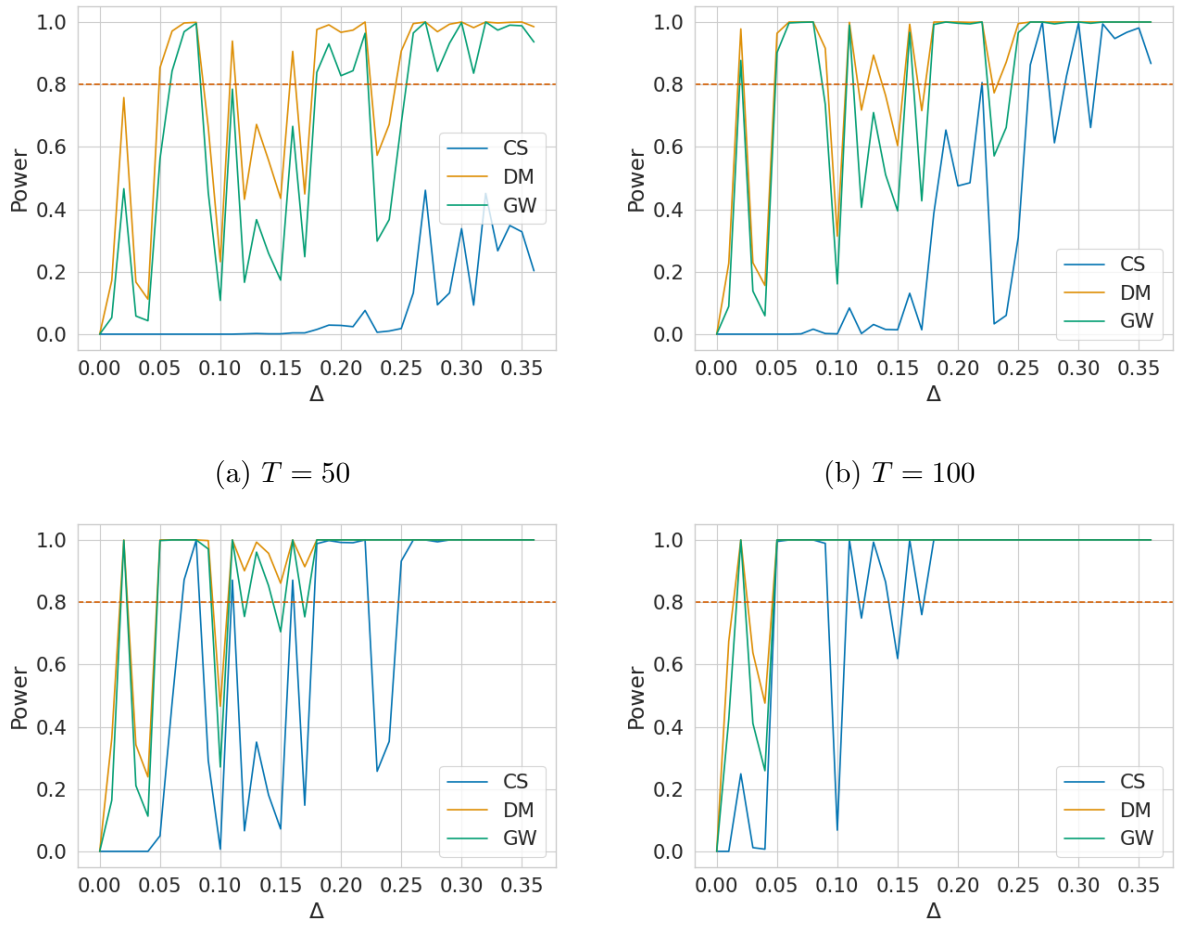
In Section 4.1, we analysed the power curves of confidence sequences with the outcome variable Y_t following Bernoulli($p = 0.7$) distribution. To exclude the possibility that the distribution influences our results we report the power curves of confidence sequences with $Y_t \sim$ Bernoulli($p = 0.7$) in Figure 14.

We notice that the power curves differ from those in Figure 2 in Section 4.1 in the following aspect. The power curves in Figure 14 are more volatile than those in Figure 2. Similar to Figure 4.1, the volatility of the power curves decreases in the sample size T.

The power curves tell a similar story as the power curves in Figure 14. The DM test reports the highest power followed by the GW test and confidence sequences.

D Programming Code

The zip file accompanying this thesis contains the files needed to generate the plots. The folder *Scripts Empirical Analysis* contains six scripts that contain the code for the sequential comparison of Section 5, two scripts for each comparison. The scripts *emp_cons_spf.py*



(a) $T = 50$

(b) $T = 100$

(c) $T = 200$

(d) $T = 500$

Figure 14: Power of confidence sequences (CS), DM and GW test for a given score differential Δ for various sample sizes.

and *emp_I_cons_spf.py* sequentially compare the growth and inflation forecasts constructed from the Consumer Sentiment Index and the Survey of Professional Forecasters, respectively. The scripts *emp_cons_liv.py* and *emp_I_cons_liv.py* sequentially compare the Consumer Sentiment Forecasts with those constructed from Livingston Survey. The comparison of the *professional* and *economist* forecasts is contained in *emp_spf_liv.py* and *emp_I_spf_liv.py*.

The folder *Scripts Simulation* contains the code for the simulations of Section 4. The script *sim_pow_analysis.py* contains power analysis plotted in Figure 2. The code producing Figures 3, 4 and 5 is contained in the script *sim_henzi_2.py*. The script *sim_trend.py* contains the simulation discussed in Section 4.2. The scripts *sim_henzi_1.py* and *sim_cum_misc.py* contain code producing Figures 7 and 8, respectively.

The folder *plots* contains all the plots produced and the folder *lib* contains frequently used function. The code was created using Python 3.7.