Real-Time Detection of Regimes of Predictability in the U.S. Equity Premium*

David I. Harvey[a], Stephen J. Leybourne[a], Robert Sollis[b] and A.M. Robert Taylor[c]

[a]School of Economics, University of Nottingham

[b]Newcastle University Business School [c]Essex Business School, University of Essex

March 9, 2020

Abstract

We propose new real-time monitoring procedures for the emergence of end-of-sample predictive regimes using sequential implementations of standard (heteroskedasticity-robust) regression t-statistics for predictability applied over relatively short time periods. The procedures we develop can also be used for detecting historical regimes of temporary predictability. Our proposed methods are robust to both the degree of persistence and endogeneity of the regressors in the predictive regression and to certain forms of heteroskedasticity in the shocks. We discuss how the monitoring procedures can be designed such that their false positive rate can be set by the practitioner at the start of the monitoring period using detection rules based on information obtained from the data in a training period. We use these new monitoring procedures to investigate the presence of regime changes in the predictability of the U.S. equity premium at the one-month horizon by traditional macroeconomic and financial variables, and by binary technical analysis indicators. Our results suggest that the one-month ahead equity premium has *temporarily* been predictable, displaying so-called 'pockets of predictability', and that these episodes of predictability could have been detected in real-time by practitioners using our proposed methodology.

Keywords: Predictive regression; persistence; temporary predictability; subsampling; U.S. equity premium.

JEL Classification: C12, C32.

1 Introduction

A large body of empirical research has been undertaken investigating stock return predictability, with a wide array of financial and macroeconomic variables considered as putative predictors

*We are grateful to three anonymous referees and the Editor, Barbara Rossi, for their helpful and constructive comments on earlier versions of this paper. We also thank participants at the EC^2 conference held in Rome in December 2018, the 50 Years of Econometrics conference held at the University of Kent in September 2018, and the Econometric Society Australasian Meeting held in Perth in July 2019 for helpful comments. Taylor gratefully acknowledges financial support provided by the Economic and Social Research Council of the United Kingdom under research grant ES/R00496X/1. Correspondence to: Robert Taylor, Essex Business School, University of Essex, Wivenhoe Park, Colchester, CO4 3SQ, United Kingdom. Email: rtaylor@essex.ac.uk.

for returns. These include valuation ratios such as the dividend-price ratio, earnings-price ratio, book-to-market ratio, various interest rates and interest rate spreads, and macroeconomic variables including inflation and industrial production; see, for example, Fama (1981), Keim and Stambaugh (1986), Campbell (1987), Campbell and Shiller (1988a,1988b), Fama and French (1988,1989) and Fama (1990). Focusing on the in-sample predictability of U.S. stock index returns these studies find relatively weak statistical evidence on predictability over short horizons, but as the forecasting horizon increases the evidence on predictability strengthens, and for longer horizons is strongly statistically significant. Finding that stock returns are predictable using financial ratios and macroeconomic variables does not necessarily mean that stock markets are inefficient. From a linearisation of the standard present value model, if the dividend-price ratio for a stock varies over time then it *must* forecast either the dividend growth rate or returns, to some extent; see, *inter alia*, Campbell and Shiller (1988a,1988b) and Cochrane (2008). More generally, if a stock market is efficient then the expected excess return for the relevant stocks might be predictable using a variety of financial and macroeconomic variables if investors' risk premia are time-varying and correlated with the business cycle.

Although consistent with orthodox financial theory, it has been argued there are statistical reasons to suspect that the strong support for predictability obtained in earlier studies could be spurious. Nelson and Kim (1993) and Stambaugh (1999) show that high persistence predictors lead to biased coefficients in predictive regressions if the innovations driving the predictors are correlated with returns, as is known to be the case for many of the popular macroeconomic and financial predictors used. Goyal and Welch (2003) show that the persistence of dividend-based valuation ratios increased significantly over the typical sample periods used in empirical studies of predictability, and argue that as a consequence out-of-sample predictions using these variables are no better than those from a no-change strategy. When estimation and inference techniques are used that take account of the high degree of persistence of the typical financial and macroeconomic predictors used, the statistical evidence of short- and long-horizon predictability is considerably weaker and in some cases disappears completely; see, *inter alia*, Ang and Bekaert (2007), Boudoukh, *et al.* (2007), Welch and Goyal (2008) and Breitung and Demetrescu (2015).

The vast majority of empirical studies of stock market predictability are based on the assumption of a constant parameter predictive regression model. However, there are several reasons to suspect that if stock returns are predictable, then it is likely to be a time-varying phenomenon; for example, significant changes in monetary policy and financial regulations could lead to shifts in the relationship between macroeconomic variables and the fundamental value of stocks, via the impact of these changes on economic growth and the growth rates of earnings and dividends. A growing body of empirical evidence is also supportive of this view. For example, Henkel *et al.* (2011) find that return predictability in the stock market appears to be closely linked to economic recessions with dividend yield and term structure variables displaying predictive power only during recessions. Timmermann (2008) argues that for most time periods stock returns are not predictable but that there are 'pockets in time' where evidence of local predictability is seen. In particular, if predictability exists as a result of market inefficiency rather than because of time-varying risk premia, then rational investors will attempt to exploit its presence to earn abnormal profits. Assuming that a large enough proportion of the total number of investors are rational, this behaviour will eventually cause the predictive power of the relevant predictor to be eliminated. If a variable begins to have predictive power for stock returns then a short window of predictability might exist before investors learn about the new relationship between that variable and returns, but it will eventually disappear; see, in particular, Paye and Timmermann (2006) and Timmermann (2008). It therefore seems reasonable to consider the possibility that the predictive relationship might change over time, so that over a long span of data one may observe some, possibly relatively short, windows of time during which predictability occurs. In such cases, standard predictability tests based on the full sample of available data will have very low power to detect these short-lived predictive episodes.

Several empirical studies find evidence suggesting that parameter instability is a feature of return prediction models. Lettau and Ludvigsson (2001) find instability in the predictive ability of the dividend and earnings yield in the second half of the 1990s. Goyal and Welch (2003) and Ang and Bekaert (2007) find instability in prediction models for U.S. stock returns based on the dividend yield in the 1990s. Paye and Timmermann (2006) undertake a comprehensive analysis of prediction model instability for international stock market indices using the Bai and Perron (1998,2003) structural change tests. They find evidence of structural breaks for many of the countries considered, arguing that "Empirical evidence of predictability is not uniform over time and is concentrated in certain periods." *op.cit.* p.312. They find some evidence of a common break for the U.S. and U.K. in 1974-1975, and for European stock markets linked to the introduction of the European Monetary System in 1979. However, it is important to stress that conventional parameter instability tests such as Chow tests and Bai-Perron tests are not valid for use with highly persistent, endogenous predictors. Indeed, Paye and Timmermann (2006) use Monte Carlo simulations to show that in such cases this can cause substantial size inflation in the Bai-Perron tests coupled with a lack of power because of the large amount of noise typically present in predictive regression models. Moreover, traditional regression t-tests for predictability and structural break tests are an *ex post* tool for detecting the statistical significance of regressors and structural breaks in a historical sample of data. They are less useful in monitoring for change in real time because their repeated application in prediction models can lead to size distortions (with the probability of at least one of the tests rejecting tending to unity as the number of tests in the sequence increases) and, as a consequence, spurious evidence of in-sample predictive ability; see Inoue and Rossi (2005) for a detailed discussion of this problem in relation to t-tests.

Motivated by this, we develop new statistical monitoring techniques, specifically designed to avoid the spurious detection problems discussed in Inoue and Rossi (2005). We use these methods to monitor the stability of predictive regression models for the U.S. equity premium. As putative predictors we consider various commonly used traditional macroeconomic and financial variables as well as a range of technical analysis rules where only price or volume data is used to predict returns. In an early paper in this direction, Brock *et al.* (1992) study the ability of moving average and trading range break trading rules to predict the Dow Jones Industrial Average (DJIA) index using daily data from 1897 through to 1986, finding strongly significant evidence that the trading strategies generated abnormal returns that cannot be explained by serial correlation or conditional heteroskedasticity in the returns. Sullivan *et al.* (1999) analyse a longer data sample on the DJIA, and find that the rules employed by Brock *et al.* (1992) were unable to identify profitable trading strategies for the period 1987-1996, although there was some evidence that they managed to do so prior to this period. Hudson *et al.* (1996) undertake a similar analysis to Brock *et al.* (1992) for UK stock index returns and find that although the rules examined do have predictive power, their use would not enable investors to make abnormal returns once trading transaction costs are accounted for. More recently Neely *et al.* (2014) have investigated the in-sample and out-of-sample predictive power of binary technical analysis indicators in a predictive regression-based context. Indicators are constructed from moving-average rules, momentum rules, and on-balance volume rules. They find the indicators have predictive power that emulates that of the traditional financial and macroeconomic variables. They also show that combining information from technical analysis indicators and macroeconomic variables significantly improves equity risk premium forecasts versus using either type in isolation.

The real-time monitoring procedures we propose are designed with the aim of detecting, as soon as possible after their inception, relatively short windows of predictability arising from shifts in the parameter on the predictor variable in the predictive regression. The presence of short pockets of predictability amongst long periods of no predictability in U.S. stock returns has recently been documented by Farmer *et al.* (2019), using nonparametric methods and employing an R^2-type statistic to measure predictability strength. Our analysis is also related to work by Dangl and Halling (2012) who use Bayesian methods to investigate gradual changes in return predictability. Although our procedures are designed to detect short regimes of predictability when the regime change is discrete, they can also be used to detect predictive regimes when the regime change is gradual and we investigate this issue with Monte Carlo simulations. Our focus is on the real-time detection of such regimes, but the methods we use can also be used for an historical analysis of the stability of predictive regression models. Our detection procedures are based around the sequential application of simple heteroskedasticity-robust regression t-statistics for the significance of the predictor variable calculated over a subsample of fixed length m. When used as simple one-shot tests these statistics can be compared with estimated critical values obtained from a training period using the subsampling-like method of Andrews (2003) and Andrews and Kim (2006). It is important to notice that these resulting one-shot tests will be able to detect general structural change in the slope parameter on the predictor variable (in that particular subsample, relative to the rest of the sample) not just a change to predictability within the given subsample. This is because a rejection will occur where the estimated slope coefficient on the predictor differs significantly between the subsample over which the one-shot test is based and the subsamples used in the critical value generation. Based on the arguments above and the work of, among others, Paye and Timmermann (2006) and Timmermann (2008), it seems reasonable to focus attention on the null model of no predictive relationship, such that structural change where it should occur is between no predictability and a short window of predictability. It is this interpretation that we will focus on in motivating and outlining our procedure. In our application to U.S. equity data we first apply standard predictability tests to the full data sets (and indeed the training periods used to obtain the estimated critical value) to check for any evidence of sustained predictability in those samples.

Our approach is based on the sequential application of these one-shot subsample test statistics commencing from a given start date. Because this is based on a sequence of subsample statistics, we need to avoid the issue of spurious detections highlighted by Inoue and Rossi (2005) by allowing the practitioner to control the overall false positive detection rate for the resulting procedure. To this end, we suggest two possible detection procedures, both of which are based on information obtained from the data in the training period. Applied using end-of-sample forms of the subsample predictability tests, both of these approaches can be used to provide a real-time monitoring procedure for the emergence of a regime of predictive ability of a regressor for returns data. The first procedure involves comparing the sequence of statistics in the monitoring period with the extremal value of the statistic (either the most negative, most positive or largest in absolute value, as appropriate to the alternative hypothesis being tested) within the training period. A predictability regime is signalled if one obtains an outcome of the predictability statistic which exceeds this extreme value from the training period. Under the second procedure we discuss, a predictability regime is deemed to have occurred if and when the number of consecutive rejections (at a given marginal significance level using a critical value estimated by subsampling from the training period) by the one-shot tests observed in the monitoring period exceeds the longest run of such rejections in the training period. Both procedures can also be used to form estimates of the locations of the signalled predictive regimes.

The remainder of the paper is organised as follows. Section 2 outlines the time-varying predictive regression model forming the basis for our analysis. Section 3 details our proposed approach to detecting windows of predictability and for dating any predictive regimes signalled, showing how to implement real-time detection procedures whose false positive detection rates can be controlled in practical applications. Section 4 reports the results from Monte Carlo simulations to investigate the finite sample behaviour of our proposed procedures. Section 5 presents an applied investigation into the predictability of the one month-ahead equity premium on the S&P Composite index. Section 6 concludes. A supplementary appendix contains a proof of Proposition 1 as well as additional Monte Carlo results (these results are summarised in section 4.2) and additional material relating to the empirical application discussed in section 5.

2 The Predictive Regime Model

We assume a relationship between the equity premium, y_t, and a single predictor variable[1] x_t that can be described by the following data generation process (DGP),

$$y_t = \mu_y + \sum_{j=1}^{n} \beta_j d_t(e_j, m_j) x_{t-1} + \epsilon_{y,t}, \quad t = 1, ..., T \tag{1}$$

where the (putative) predictor is generated by

$$\begin{aligned} x_t &= \mu_x + s_{x,t}, \quad t = 0, ..., T && (2) \\ s_{x,t} &= \rho s_{x,t-1} + \epsilon_{x,t}, \quad t = 1, ..., T && (3) \end{aligned}$$

with $s_{x,0} = 0$ and where $d_t(e_j, m_j)$ is a dummy variable defined such that $d_t(e_j, m_j)$ takes the value 1 for $m_j > 0$ consecutive values of t, ending with $t = e_j$. The innovation vector $\epsilon_t := [\epsilon_{y,t}, \epsilon_{x,t}]'$, where the notation "$x := y$" denotes that x is defined by y, is assumed to be a strictly stationary and uncorrelated mean zero process with unconditional covariance matrix given by

$$E(\epsilon_t \epsilon_t') = \begin{bmatrix} \sigma_y^2 & r_{xy}\sigma_y\sigma_x \\ r_{xy}\sigma_y\sigma_x & \sigma_x^2 \end{bmatrix}$$

where r_{xy}, $|r_{xy}| < 1$, is the correlation between $\epsilon_{y,t}$ and $\epsilon_{x,t}$. Notice that our assumption on ϵ_t allows for the presence of conditional heteroskedasticity, such as GARCH or stationary autoregressive stochastic volatility, in both $\epsilon_{y,t}$ and $\epsilon_{x,t}$.

In the context of (1), if $\beta_j \neq 0$, then we have a *predictive regime* of y_t by x_{t-1} of length m_j observations running from $t = e_j - m_j + 1$ through to $t = e_j$. The model in (1) allows for $n \geq 0$ such predictive regimes. Consistent with the discussion in the introduction and Paye and Timmermann (2006) and Timmermann (2008), we have in mind scenarios where such regimes are relatively scarce and short-lived so that both the number of predictive regimes, n, and their durations, m_j, $j = 1, ..., n$, are taken to be small relative to the sample size, T. We assume $e_j < e_{j+1} - m_{j+1}$ such that the regimes where predictability holds are ordered (i.e. $d_t(e_1, m_1)$ is the earliest regime) and non-overlapping. Our proposed predictive regime detection procedure will consider the quantities e_j and m_j which delimit the start and end dates of the predictive regimes, and the number of regimes, n, to be unknown to the practitioner. Outside of these n

[1] For lucidity, we outline our procedure for the case of a single predictor. Our approach can be extended to the case where multiple predictors feature in (1). Here individual subsample t-statistics, of the form discussed in section 3.1, associated with each of the predictor variables could be considered along with multi-parameter heteroskedasticity-robust regression F-statistics. Consideration would need to be given to the appropriate statistics and decision rules to adopt, and to the usual issues surrounding multiple (significance) testing. Moreover, although we focus on the case where a constant term is included in both (1) and (2), our approach is also valid for a more general deterministic component, such as a polynomial deterministic trend, appearing in both components provided it is included in the test regression in (4) and the t-statistic, $\tau_{e,m}$, in (5) is commensurately redefined.

predictive regimes the slope parameter in (1) is zero and the DGP is such that $y_t = \mu_y + \epsilon_{y,t}$ and, hence, y_t is unpredictable (in mean) due to the $\epsilon_{y,t}$ being serially uncorrelated (a standard assumption in this literature). Where $n = 0$ in (1), y_t is unpredictable at all time periods.

As is standard in this literature, we have adopted an AR(1) specification for $s_{x,t}$, and hence for x_t, in (3). As we will discuss in section 3, the predictive regime detection procedures we propose in this paper can be applied regardless of whether the autoregressive root, ρ, in (3) is such that $\rho = 1$ (a unit root predictor) or $|\rho| < 1$ (a stationary predictor). Moreover, ρ is also allowed to be T-dependent such as occurs, for example, in cases where the predictor is strongly persistent displaying either local or moderate deviations from a unit root; for full sample predictability tests directed at the latter, see Kostakis *et al.* (2015). The AR(1) specification in (3) is not critical for our analysis, and it could be generalised to allow $\epsilon_{x,t}$ to be a weakly autocorrelated process without affecting the validity of our proposed procedures; see Remark 3 in section 3.2.1.

In what follows, to facilitate our later analysis of real-time monitoring for the emergence of predictive regimes, we make a distinction between the end of the monitoring period, which we denote by $t = E$, and the notional future end of the DGP for y_t, that is $t = T$, such that $E \leq T$.

3 Predictive Regime Detection

3.1 Subsample Regression t-statistics

We are interested in detecting the presence of a predictive regime for y_t in real time and propose a way of doing this using subsample regression t-statistics. To that end, consider first selecting a subsample of m observations running from $t = e - m + 1$ to $t = e$, where m is a fixed value (independent of the sample size, T) chosen by the user, and run the (generic) ordinary least squares regression,

$$y_t = a + bx_{t-1} + u_t, \qquad t = e - m + 1, ..., e. \tag{4}$$

We then calculate the regression t-statistic, based around a heteroskedasticity-robust variance estimate (see White, 1982), for the significance of x_{t-1} in (4); that is,

$$\tau_{e,m} := \hat{b}/(\hat{V}(\hat{b}))^{1/2} \tag{5}$$

where

$$\hat{b} \ := \ \frac{\sum_{t=e-m+1}^{e}(x_{t-1} - \bar{x}_{-1})(y_t - \bar{y})}{\sum_{t=e-m+1}^{e}(x_{t-1} - \bar{x}_{-1})^2}, \quad \hat{V}(\hat{b}) := \frac{\sum_{t=e-m+1}^{e}(x_{t-1} - \bar{x}_{-1})^2\hat{u}_t^2}{\{\sum_{t=e-m+1}^{e}(x_{t-1} - \bar{x}_{-1})^2\}^2} \tag{6}$$

$$\hat{u}_t \ := \ (y_t - \bar{y}) - \hat{b}(x_{t-1} - \bar{x}_{-1}) \tag{7}$$

$$\bar{y} \ := \ m^{-1}\sum_{t=e-m+1}^{e} y_t, \quad \bar{x}_{-1} := m^{-1}\sum_{t=e-m+1}^{e} x_{t-1}.$$

Detection of a predictive regime holding between y_t and x_{t-1} for the given subsample $t = e-m+1, ..., e$ can be based on $\tau_{e,m}$. As a particular example, suppose we have data available for $t = 1, ..., T^* + m \leq T$; a test for the presence of a predictive regime in the last m available sample observations would therefore be based on the statistic $\tau_{T^*+m,m}$. Standard regime detection tests, such as those outlined in Paye and Timmermann (2006) use asymptotic (in the sample size T) distribution theory to approximate the test's critical value, but this approximation is based on the assumption that the sample window m used in constructing the statistic is a positive fraction of T. This assumption is clearly not consistent with our aim of detecting predictive regimes of short duration. Moreover, even if we were to assume m to be a function of T, the limiting distribution of $\tau_{e,m}$ will depend on nuisance parameters in the DGP in (1)-(3); specifically, the degree of persistence of the predictor variable, x_t, and the correlation, r_{xy}, between $\epsilon_{y,t}$ and $\epsilon_{x,t}$. Without knowledge of these, valid asymptotic critical values could not be obtained.

An alternative approach, which we will consider further in the context of the detection procedure proposed in section 3.2.2, robust to the degree of persistence and endogeneity of the predictor can be based on the subsampling approach of Andrews (2003) and Andrews and Kim (2006). In the end-of-sample example above, suppose we have a sample of size $T^* + m$ and we form the predictability statistic $\tau_{T^*+m,m}$. To obtain a critical value, one uses the *training period* $t = 1, ..., T^*$, to compute the $T^* - m - 1$ analogous statistics $\{\tau_{e,m}\}$, $e = m + 1, ..., T^*$. The $(1 - \pi)$ sample quantile of these statistics is the estimated significance level-π critical value for the end-of-sample predictability test. By construction, the resulting test is (asymptotically in T) robust to nuisance parameters in (1)-(3) because the training period statistics have the same functional dependence on those nuisance parameters as $\tau_{T^*+m,m}$. This test will have non-trivial power whenever there is predictability in the last m observations, but not in the training period.

Crucially though, the discussion above relates to a one-shot predictability test. However, our goal in this paper is to develop real-time monitoring procedures for the emergence of an end-of-sample predictive regime. To that end, we will construct a sequence of $\tau_{e,m}$ statistics, of the form given in (5), calculated for each possible end-of-subsample date $e = T^* + m, ..., E$, recalling that E denotes the end of the monitoring period, a parameter set by the practitioner. The predictive regime detection procedures we propose below are based on comparing the behaviour of this sequence of statistics with corresponding sequences within the training period and will be designed such that the theoretical (i.e. large sample) false positive rate [FPR] of the procedures is known and can be properly controlled, where the FPR represents the probability of incorrectly signifying the presence of at least one predictive regime in the monitoring period.

3.2 The Detection Procedures

We now detail our predictive regime detection approaches. For transparency, these are presented in the context of upper tail testing (i.e. for predictability regimes where $\beta_j > 0$), but can be adapted to lower tailed or two tailed testing in an obvious way. We will discuss two procedures, each of which forms a decision rule for rejecting the null of no predictability in the monitoring period based on specific properties of the sequence of $\tau_{e,m}$ statistics within the given training period. The first procedure we consider will based on the largest of the $\tau_{e,m}$ statistics observed in the training period, and the second will be based on the longest run of outcomes of the $\tau_{e,m}$ statistics in the training period that exceed a given (critical) value.

For both of the procedures which follow, we define the training period as $t = 1, ..., T^*$. We assume that no predictive regime occurs within the training period; that is, $T^* < e_1 - m_1 + 1$; further discussion relating to where this assumption might be violated is given in section 3.4. In what follows we assume that T^* and E are such that $T^* := \lfloor \lambda_1 T \rfloor$, and $E := \lfloor \lambda_2 T \rfloor$, $\lfloor \cdot \rfloor$ denoting the integer part of its argument, and where $0 < \lambda_1 < \lambda_2 \leq 1$.

3.2.1 The MAX Procedure

The first detection procedure we propose, which we will denote as MAX, is based on the maximum value of the sequence of $\tau_{e,m}$ statistics taken across the training and monitoring periods (cf. Astill *et al.*, 2018). More precisely, with $\{\tau_{e,m}\}_{e=m+1}^{T^*}$ and $\{\tau_{e,m}\}_{e=T^*+m}^{E}$ constituting the statistics obtained from the training and monitoring periods, respectively, we consider a detection procedure whereby a predictive regime in the monitoring period is signalled if $\max_{e\in[T^*+m,E]} \tau_{e,m}$ exceeds $\max_{e\in[m+1,T^*]} \tau_{e,m}$; that is, the largest $\tau_{e,m}$ in the monitoring period exceeds the largest $\tau_{e,m}$ in the training period.

We now establish the theoretical (as $T \to \infty$) FPR of the MAX procedure when run out to the end of monitoring date, E. This is done by evaluating the limiting probability that $\max_{e\in[T^*+m,E]} \tau_{e,m} > \max_{e\in[m+1,T^*]} \tau_{e,m}$ under the null hypothesis that no predictability is present in the DGP. This result is now given in Proposition 1.

Proposition 1. *Let* (y_t, x_t) *be generated according to* (1)-(3) *under the conditions stated in section 2. Let the MAX decision rule be as given above. If $n = 0$, such that no predictability is present in the DGP, then as $T \to \infty$,*

$$\lim_{T\to\infty} P\left(\max_{e\in[T^*+m,E]} \tau_{e,m} > \max_{e\in[m+1,T^*]} \tau_{e,m}\right) = \alpha^* \tag{8}$$

where $\alpha^* := (\lambda_2 - \lambda_1)/\lambda_2 = \lim_{T\to\infty} \alpha$ where, for the stated choices of monitoring and training periods,

$$\alpha := \left(\frac{E - T^* - m + 1}{E - 2m + 1}\right). \tag{9}$$

Remark 1. *The result in Proposition 1 provides an expression for the theoretical FPR of the MAX decision rule; that is, the limiting probability that the maximum of the $\tau_{e,m}$ statistics in the monitoring period exceeds the maximum of the $\tau_{e,m}$ statistics in the training period in the case where no predictability occurs. This is seen to be simply the limiting value of the ratio* *formed by dividing the total number of $\tau_{e,m}$ statistics computed in the monitoring period (here $E - T^* - m + 1$) by the total number of $\tau_{e,m}$ statistics calculated in the training and monitoring periods combined (here $(E - T^* - m + 1) + (T^* - m) = E - 2m + 1$). This result holds more generally when comparing the maxima of the sequences of $\tau_{e,m}$ statistics obtained from any two disjoint subintervals of the data whose lengths are both functions of T.*

Remark 2. *The result in Proposition 1 holds regardless of the degrees of persistence and endogeneity of the regressors in the predictive regression and holds for all conditionally heteroskedastic innovations which satisfy the condition of strict stationarity. In particular, the result in Proposition 1 holds regardless of whether the putative predictor x_t in* (3) *is: weakly dependent ($|\rho| < 1$); strongly persistent ($\rho = 1 - c/T$ with the constant $c \geq 0$, where $c = 0$ yields the pure unit root case, while $c > 0$ corresponds to the local-to-unity case); or moderately persistent ($\rho = 1 - cT^{-\theta}$ with $c > 0$ and $\theta \in (0, 1)$, the moderate deviations from unity case of Kostakis et al., 2015).*

Remark 3. *As demonstrated in the proof of Proposition 1, the stated result follows using an application of Theorem 2.1 of Ferreira and Scotto (2002,p.478), with $r = s = 1$ in their notation, which applies to strictly stationary sequences of mixing random variables. To do so we establish that under the conditions given in section 2, $\{\tau_{e,m}\}$ forms a strictly stationary and $(m - 1)$ dependent sequence, the latter therefore satisfying the required mixing condition stated in Ferreira and Scotto (2002,p.476). We have assumed for simplicity that ϵ_t is serially uncorrelated which yields the $(m - 1)$ dependence result. Weakening this assumption to allow for stationary serial correlation in $\epsilon_{x,t}$ would not alter this result. It is standard in this literature to assume that $\epsilon_{y,t}$ is serially uncorrelated. However, this could be weakened to allow finite $MA(k)$, $0 \leq k < \infty$, behaviour in $\epsilon_{y,t}$ in which case $\{\tau_{e,m}\}$ would be a $(k + m - 1)$ dependent sequence but would still satisfy the required mixing condition. We cannot formally allow for unconditional heteroskedasticity in ϵ_t because $\{\tau_{e,m}\}$ would not then form a strictly stationary sequence and so we could not appeal to Theorem 2.1 of Ferreira and Scotto (2002). However, we have still based our approach on heteroskedasticity-robust t-statistics because although not exact invariant to any unconditional heteroskedasticity present (which is what would be needed as m is finite) we expect them to be considerably more robust than the corresponding t-statistics based on OLS standard errors. In section 4 we will investigate the impact of unconditional heteroskedasticity in ϵ_t on the finite sample FPRs of the procedures discussed in this section.*

For given values of T^* and m, we can use (9) to approximate the empirical FPR that would be obtained in practice for any monitoring horizon E. We observe that α is a monotonically increasing function of E as $\frac{\partial \alpha}{\partial E} = \frac{T^* - m}{(E - 2m + 1)^2} > 0$. Hence, other things being equal, the longer the monitoring period, the greater the likelihood of spuriously finding a predictive regime. To illustrate, Figure 1 graphs this approximation for the case of $T^* = 400$ and $m = 30$. So, for example, reading from Figure 1, if we wish to monitor out to $E = 680$, then the FPR will be about 0.40.

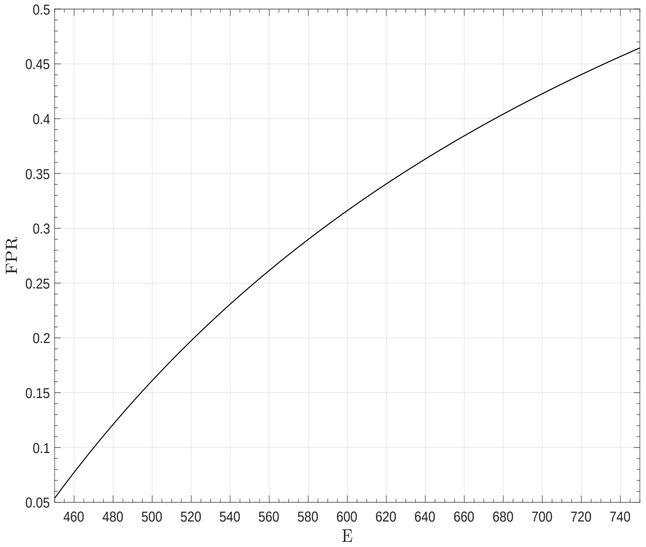


Figure 1. FPR as a function of E

We can also rearrange (9) as

$$E = \frac{T^{*} + m - 1 - \alpha(2m-1)}{1-\alpha} \tag{10}$$

which is useful if we wish to know the maximum monitoring horizon E such that the FPR for the MAX procedure is (approximately) controlled at α. For the current illustration, (10) shows that E should be chosen to be no more than about 520 for a choice of $\alpha = 0.20$ (which is also apparent from Figure 1).

3.2.2 The *SEQ* Procedure

Our second detection procedure, which we denote as SEQ, is based on comparing the length of the longest *contiguous sequence* of exceedances of some value pre-set by the practitioner by the statistics $\tau_{e,m}$ in the monitoring period, with the corresponding measure taken over the training period. An obvious choice for this threshold value, which we will adopt in what follows, would be to use a relevant marginal critical value for some significance level π for the one-shot $\tau_{e,m}$ test.[2] In doing so we will follow the subsampling approach of Andrews (2003) and Andrews and Kim (2006) and calculate an empirical critical value, denoted cv_{π} in what follows, from the training period. Recalling that the sequence of $\tau_{e,m}$ statistics that make use of data within the training period is given by $\tau_{e,m}$ for $e = m+1, ..., T^{*}$, then cv_{π} is defined such that $cv_{\pi} := \tau_{(\lfloor(1-\pi)(T^{*}-m)\rfloor)}$ where $\tau_{(j)}$, $j = 1, ..., T^{*} - m$ are the ascending order statistics of $\tau_{e,m}$, $e = m+1, ..., T^{*}$ (that is,

[2] Any sensible threshold value could in principle be used. A benefit of using such a critical value is that where the training period contains no predictive regimes each individual test in our monitoring sequence can be interpreted marginally as a test for predictability in that particular subsample. As such, it makes sense in practice to set π to a conventional significance level, e.g. $\pi = 0.05$ or $\pi = 0.10$.

$\tau_{(j+1)} > \tau_{(j)}$ for $j = 1, ..., T^* - m - 1$). Under the conditions on the DGP considered by Andrews and Kim (2006), cv_π is a consistent (as $T \to \infty$) estimate for the true π significance level critical value. However, it should be stressed that the SEQ procedure we propose does not rely on this consistency property holding on cv_π.

Based on cv_π we then consider the maximum number of contiguous values of $\tau_{e,m}$ within the training period that exceed cv_π. To this end, define $R_{\pi,e} := 1(\tau_{e,m} > cv_\pi)$, where $1(\cdot)$ denotes the indicator function, and consider the following measure over $e = L$ to $e = U$ with $U \geq L$

$$R_\pi(L, U) := (U - L + 1) \prod_{e=L}^{U} R_{\pi,e}$$

Here, when $R_\pi(L, U)$ is non-zero, its value, $U - L + 1$, represents the length of a sequence of contiguous exceedances. The maximum length of contiguous exceedances in the training period is then given by $\max_{L,U \in [m+1,T^*]} R_\pi(L, U)$. The corresponding measure for the monitoring period is given by $\max_{L,U \in [T^*+m,E]} R_\pi(L, U)$. Our proposed SEQ procedure is then to signal a predictive regime in the monitoring period if $\max_{L,U \in [T^*+m,E]} R_\pi(L, U) > \max_{L,U \in [m+1,T^*]} R_\pi(L, U)$. Paralleling the result in Proposition 1, when there is no predictability in the training or monitoring periods we conjecture that

$$\lim_{T\to\infty} \Pr\left(\max_{L,U \in [T^*+m,E]} R_\pi(L, U) > \max_{L,U \in [m+1,T^*]} R_\pi(L, U)\right) \leq \alpha^* \tag{11}$$

where α^* is as defined in Proposition 1. Notice here that, in contrast to the result for the MAX monitoring procedure where the large sample FPR when monitored up to E is exactly α^*, the corresponding quantity for the SEQ procedure is *bounded* by α^*. This arises because $\max R(L, U)$ can only assume integer values, so there is a non-zero probability of a tied value in the training and monitoring periods, even asymptotically. Hence the strict equality obtained for the MAX procedure from Proposition 1 is replaced by the weak inequality in (11). The (approximate) relationship in (10) can also still be considered to hold, but interpreted to be the maximum monitoring horizon E such that the FPR for the SEQ procedure is bounded by α.[3]

It will be convenient to denote the training period maximum length of contiguous exceedances, $\max_{L,U \in [m+1,T^*]} R_\pi(L, U)$, as l_π. Notice that the first time period at which it would be possible for SEQ to signal a predictive regime is $t = T^* + m + l_\pi$, because this is the first occasion where $R_\pi(L, U)$ in the monitoring period can exceed l_π. In contrast, it is possible for the MAX

[3] We are unable to provide a formal proof of the result of (11), hence our conjecture on the basis of extant, but much more limited theoretical results. A formal proof would be extremely involved, if even tractable, given the complexity of the arguments needed in Ferreira and Scotto (2002) to establish theoretical results relating to the much simpler case of subsample maxima. However, this conjecture is not without foundation. We have also conducted extensive Monte Carlo simulation experiments that appear to support it. Furthermore, these simulation results reveal that the empirical FPR of the SEQ procedure is always below but very close to α, implying that the probability of tied values in the training and monitoring periods is very small.

procedure to signal a predictive regime as early as $t = T^* + m$. However, we can control l_π via the choice of π. The larger is π then the smaller is cv_π so we would naturally expect the larger is l_π. This relationship is important as choosing a large value of π might lead to what is considered an unacceptable delay before being able to detect a predictive regime. This is not a consideration with MAX, however. In fact, MAX can be thought of as an extreme case of SEQ where we choose $cv_\pi = \max_{e \in [m+1, T^*]} \tau_{e,m}$ (the smallest value of cv_π such that $\pi = 0$) and hence $l_\pi = 0$.

3.3 Dating of Predictive Regimes

In a real-time monitoring context, if the procedure signalled the presence of a predictive episode at time $E^* \leq E$ then the monitoring procedure would of course terminate at that point given that the procedure would have signalled the presence of a predictive regime at that time. However, one could also consider continuing the monitoring procedure up until E. It is therefore possible for both of our proposed MAX and SEQ procedures to detect more than one predictive regime before the notional end-of-monitoring date, E.

Although our focus on this paper is on real-time detection we can, where at least one predictive regime has been signalled by one of our procedures when run out until the end of the monitoring period, E, provide approximate dates for the location of these predictive regime(s). This should be viewed more as a historical dating exercise rather than something that would be done in the context of a real-time monitoring procedure. Detailing this first in the context of the MAX procedure, for $e = T^* + m, ..., E$ define $R_{0,e} := 1(\tau_{e,m} > \max_{s \in [m+1, T^*]} \tau_{s,m})$. Next, let D denote an $E \times 1$ vector of zeros, and set $D_e = 1$ whenever $R_{0,e} = 1$. Now suppose that D has h consecutive 1s in positions $e = j, ..., j + h - 1$ where j is the earliest date for which $R_{0,j} = 1$. Here $R_{0,j}$ is based on data over the period $j - m + 1, ..., j$, so we might therefore consider $j - m + 1$ to represent a feasible start date for the first predictive regime. With $R_{0,j+h-1}$ representing the final exceedance in D, and this being based on data over the period $j - m + h, ..., j + h - 1$, we might similarly consider $j + h - 1$ to represent a feasible end date for this predictive regime. By this categorisation, then, the predictive regime covers the contiguous set of dates $j - m + 1, ...,$ $j + h - 1$. In some sense, this set of dates is liberal, or *weak*, in that it is possible that the predictive regime started after $j - m + 1$ and ended before $j + h - 1$; for example, only the later data used in $R_{0,j}$ may be responsible for triggering that exceedance, and only the earlier data used in $R_{0,j+h-1}$ responsible for triggering that exceedance. We might therefore consider an alternative dating approach where the predictive regime is characterised by the subset of dates for which every time that date is present in the subsample of data being tested, an exceedance is obtained. This subset, which we will refer to as *strong*, is the contiguous set of dates $j, ..., j - m + h$; notice that if $h \leq m - 1$, the strong set will be empty. A second predictive regime is deemed to exist if $R_{0,j+h} = 0$ but $R_{0,j+h+s} = 1$ for some $s \geq 1$, and weak/strong dates for the second regime can be determined in the same manner as for the first regime. This extends to more than two regimes in an obvious way. In situations where more than one predictive regime has been detected, it is possible that weak dates associated with consecutive regimes can overlap, although this possibility cannot arise with the strong dates.

For the SEQ procedure, the dating method follows the same process as for the MAX procedure, but with the non-zero elements of the D vector defined according to the following: for $e = T^* + m + l_\pi, ..., E$, if $\prod_{k=e-l_\pi}^{e} R_{\pi,k} = 1$, set $D_{e-l_\pi}, ..., D_e$ to 1. That is, for all end-of-window dates e that form part of a contiguous run of at least $l_\pi + 1$ exceedances $R_{\pi,e}$, we set the eth element of D to one. The weak and strong dates can then be categorised in exactly the same way as for the MAX, based on the $R_{\pi,e}$ exceedances involved in the D vector.

3.4 Additional Discussion

We conclude this section with some observations, which apply in equal part to the MAX and SEQ procedures.

1. Suppose now that, in contradistinction to our maintained assumption so far, one or more predictive regimes in (1) are present within the chosen training period. Provided such regimes are of finite length and finite in number, then the asymptotic (in T and T^*) properties of the MAX and SEQ procedures are unaffected by this. For a finite length training period, if predictability regimes existed within it, we would expect both $\max_{e\in[m+1,T^*]} \tau_{e,m}$ and l_π to be increased relative to the case where no predictability is present in the training period, other things being equal. We might therefore anticipate some reduction in the ability of our procedures to detect genuine predictive regimes present in the monitoring period. We will explore the impact on our proposed procedures of a predictive regime holding in the training period as part of our Monte Carlo simulation study in section 4.

2. Although not consistent with the interpretation we are placing on the DGP in (1), as discussed in the introduction it is possible in practice that the training period could potentially exhibit predictability throughout its duration, or a large part of its duration. In this case, an upper tail rejection arising from MAX or SEQ in the monitoring period should be taken to indicate a statistically significant increase in the magnitude of the slope parameter on x_{t-1} (and, hence, in the strength of the predictability of y_t by x_{t-1}) *vis-à-vis* its value in the training period. In practical applications, we therefore recommend prior application of standard full-sample predictability tests to the training period to investigate whether the assumption of no predictability holds in the training period, and this will be done in the empirical data analysis undertaken in section 5.

3. Our discussion thus far has implicitly assumed that the training period runs from the earliest available time period in the dataset to the point immediately before the desired start of monitoring. This essentially makes the training period as large as possible, which

ensures that, through the role of T^* in (9) and (10), the FPR is as small as possible for a given E, or, equivalently, E is as large as possible for a given FPR. In cases where a very long history of data is available, it may be prudent to use only relatively recent data, to avoid including historical predictive regimes in the training period. In practice, such regimes might be detected by prior pre-testing, an approach we adopt in the empirical application in section 5. Furthermore, we have so far focused, for simplicity, on the case where there is no separation between the data period used for the training period and the data used for monitoring, with the former spanning $t = 1, ..., T^*$ and the latter starting at $t = T^* + 1$. More generally, the last time period included in the training period could be $T^* - k$ for some $k > 0$, allowing for a separation between the training period and the start of the monitoring period. This might be relevant in cases where a predictive regime was thought to have occurred towards the end of the training period, so that the training period could be redefined to exclude this regime. As noted in Remark 1, an analogous result to Proposition 1 also holds here and the expressions for α and E in (9) and (10) in this case become, respectively,

$$\alpha = \frac{E - T^* - m + 1}{E - 2m + 1 - k} \qquad \text{and} \qquad E = \frac{T^* + m - 1 - \alpha(2m - 1 + k)}{1 - \alpha}.$$

4 Finite Sample Properties of the Monitoring Procedures

We now report the results from four Monte Carlo simulation experiments designed to study the finite sample properties of our MAX and SEQ procedures. These investigate the FPRs of the two procedures and their power to detect a predictive regime of given length. Extensive additional simulations were also undertaken to study the detection power of MAX and SEQ as a function of m_1 (the length of the predictive regime in the DGP), and to study the robustness of our procedures to different error term assumptions, patterns of heteroskedasticity, to higher-order autocorrelation in the predictor, and to gradual regime change. We present these additional results in an on-line supplementary appendix and briefly discuss the key findings in section 4.2.[4].

In all of the experiments we generated the simulation data according to the DGP given by (1)-(3) and set $\mu_y = \mu_x = 0$ (without loss of generality) using negatively correlated error terms with $r_{x,y} = -0.90$.[5] For the four sets of experiments reported in the main text we generate $\epsilon_{y,t} \sim N(0, 1)$, $\epsilon_{x,t} \sim N(0, 1)$. All of the simulation experiments and the empirical application in section 5 employ the upper-tailed version of our procedure.[6] In each simulation experiment

[4] The on-line supplementary appendix is available from `www.sites.google.com/view/pr-supplementary`, which also contains the data and MATLAB code used for the paper.

[5] In predictive regression models for the equity premium employing valuation ratios as predictors (e.g. the dividend-price ratio, earnings-price ratio) the relevant error terms are strongly negatively correlated, hence our choice of $r_{x,y} = -0.90$.

[6] For the majority of the macroeconomic and financial variables and for all of the technical analysis indicators

the sample period when monitoring starts ($T^* + m$) is the same as in the empirical application, $T^* + m = 302$, and for the main experiments, $m = 30$.[7] All of the experiments are undertaken using MATLAB, employing the Mersenne Twister random number generator function and 10,000 replications.

The first set of experiments studies the power of MAX and SEQ to detect a single predictive regime as a function of $\beta_1 = \{0.05, 0.10, ..., 0.45, 0.50\}$ for $\rho = \{0.965, 0.995\}$, setting $\pi = 0.10$.[8] When $\beta_1 = 0$ (so that $n = 0$ and, hence, there are no predictive regimes in the data) the detection frequency obtained from the simulations is equivalent to an empirical FPR and we also report simulation results for this case. In this first set of experiments we assume a short monitoring period that ends at $E = 327$, which, given the values used for T^* and m, is consistent with $\alpha = 0.10$ (this can be verified using (9)). Therefore when $\beta_1 = 0$, the empirical FPR obtained for each procedure should be approximately equal to 0.10. If a predictive regime does occur during the monitoring period, then the power of our procedures to detect its presence will depend not only on how long the relevant predictive regime continues for (m_1) and its strength (measured by the magnitude of β_1), but also on when the predictive regime occurs relative to the start of monitoring. To investigate this issue in more detail, separate results are computed for five different predictive regime start dates: (a) $t = 287$ (15 observations before the start of monitoring), (b) $t = 297$ (5 observations before the start of monitoring), (c) $t = 302$ (at the same time as the start of monitoring), (d) $t = 307$ (5 observations after the start of monitoring), (e) $t = 317$ (15 observations after the start of monitoring). In each case the length of the predictive regime in the DGP is set to $m_1 = 30$.[9]

In empirical applications, whilst there might be a particular reason for favouring a short monitoring period, for predictive regimes that start towards the end of a short monitoring period the power of our procedure to detect their presence could be significantly improved if we monitor for a longer period of time. To investigate this issue in more detail, in the second set of experiments we repeat the first set of experiments employing the same simulation DGP and predictive regime dates, but extending the monitoring period to $E = 361$ which is consistent with $\alpha = 0.20$. Hence the empirical FPR obtained from the simulations in this case (when $\beta_1 = 0$) should be approximately equal to 0.20.

used in the empirical application in section 5 financial theory suggests a positive relationship with the equity premium. For those of the macroeconomic and financial variables where financial theory suggests a negative relationship with the equity premium (e.g. interest rates) we use $-x_{t-1}$ rather than x_{t-1} when testing for a predictive regime so that an upper-tailed test is applicable. This is consistent with recent research on detecting equity premium predictability using orthodox t-tests (e.g. Campbell and Thompson, 2008; Neely *et al.*, 2014).

[7] The data sample used for the equity premium application below is monthly covering the period December 1974 to December 2015 ($T = 493$). In the application we monitor from January 2000 (hence $T^* + m = 302$). In addition to $m = 30$, in the empirical application results are also computed for $m = 15$ and $m = 60$.

[8] This range of values for ρ and β_1 was chosen following a preliminary analysis of the data used for the empirical application in section 5. Typically when AR(1) models are estimated for the traditional predictors used in section 6 (e.g. the valuation ratios), the AR(1) coefficient estimates lie in the range 0.965-0.998.

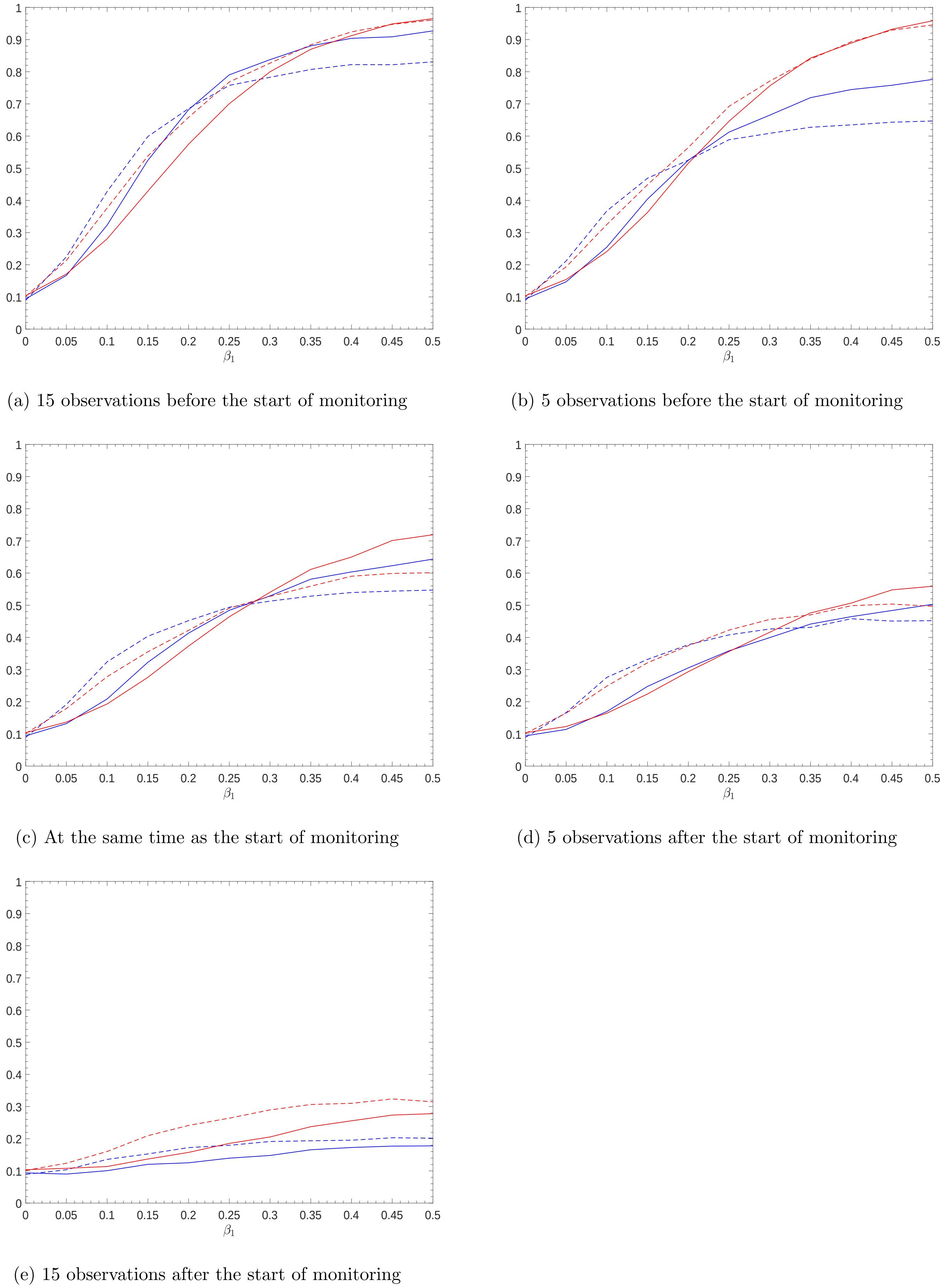
[9] Therefore in these experiments $m = m_1$. In the additional simulation experiments discussed in section 4.2, we investigate the performance of our monitoring procedure when the values of m and m_1 differ.

The first two sets of experiments assume no predictability in the training period. As discussed in section 3.4, our procedure can still be used for detecting predictive regimes during the monitoring period if predictability exists during the training period, although the FPR and power of the procedure could be affected. If our procedure is applied to data where a regime of positive predictability exists in the DGP during the training period, both the largest value of $\tau_{e,m}$ over the training period, and the longest contiguous sequence of right-tailed $\tau_{e,m}$ exceedances over the training period, are likely to be larger than the values obtained if the DGP had contained no predictability over the training period but was otherwise identical. It follows straightforwardly in this case that the power of our procedures to detect a predictive regime over the monitoring period (and also the empirical FPRs) will be reduced relative to the case of no predictability over the training period.

The third and fourth sets of experiments investigate this issue in more detail. In these experiments we repeat the first two sets of experiments again using the DGP given by (1)-(3), but in addition to the original predictive regime at locations (a)-(e), an earlier predictive regime is imposed in the DGP during the relevant training periods. Specifically, the full DGP for each set of experiments contains two predictive regimes (i.e. we set $n = 2$ in (1)), where the first predictive regime is set to occur during the training period at $t = \lfloor T^*/2 \rfloor + 1$, and we set $m_1 = 15$ and $\beta_1 = 0.25$ (hence the associated predictive regime in the training period continues for 15 observations). The second predictive regime mirrors the original predictive regime in the first two sets of experiments. The length of this second regime, m_2, and the strength of the predictability, β_2, are set to the same values as the relevant parameters in the first two sets of experiments (m_1 and β_1, respectively). Note that in the third and fourth sets of experiments the predictive regime in the training period is relatively short (being half the length of the predictive regime in the monitoring period for first two sets of experiments). It is particularly important to assess the finite sample performance of our procedures when there is a short predictive regime in the training period, since short predictive regimes are more difficult to identify than long predictive regimes. If a long predictive regime exists over the initial training period chosen by a researcher using our procedures, then it is more likely that the researcher would be aware of its presence (e.g. via a preliminary analysis of the data).

4.1 Main Results

The results from the first set of experiments are given in Figure 2 which, as with Figures 3-5, graphs the empirical frequencies with which at least one predictive regime is signalled by our monitoring procedures MAX (solid and dotted red lines) and SEQ (solid and dotted blue lines) when run across the whole monitoring period under consideration. Recall that the end of the monitoring period for the set of experiments relating to Figure 2 is chosen using (9) to be such that $\alpha = 0.10$. Therefore, when $\beta_1 = 0$ we would expect the simulated predictive regime detection



(a) 15 observations before the start of monitoring

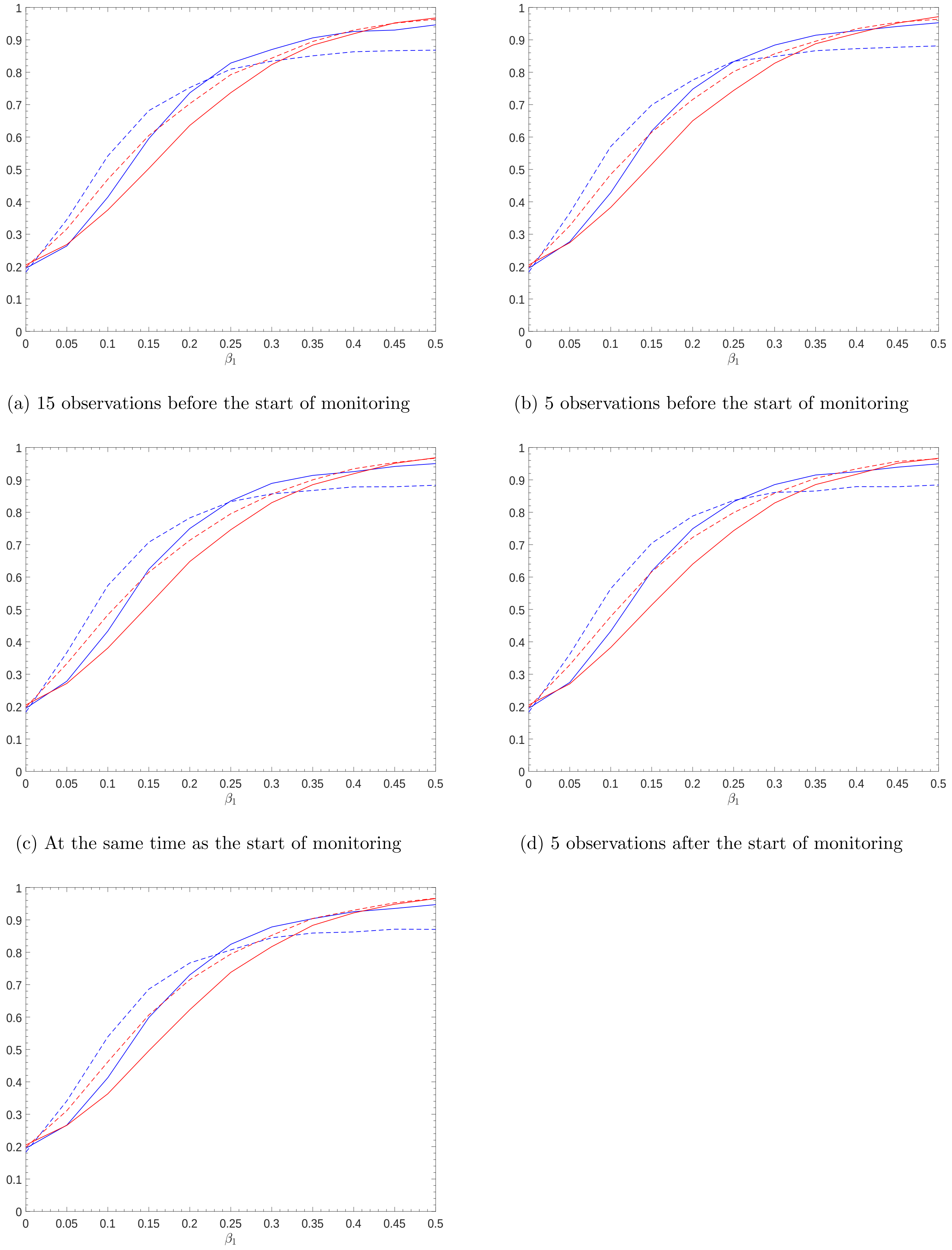
(b) 5 observations before the start of monitoring

(c) At the same time as the start of monitoring

(d) 5 observations after the start of monitoring

(e) 15 observations after the start of monitoring

Figure 2. Predictive regime detection frequency as a function of β_1 for different values of ρ: $T^* + m = 302$, $E = 327$, $m_1 = 30$, $m = 30$; MAX, $\rho = 0.965$, —; MAX, $\rho = 0.995$, - - -; SEQ, $\rho = 0.965$, —; SEQ, $\rho = 0.995$, - - -



(a) 15 observations before the start of monitoring

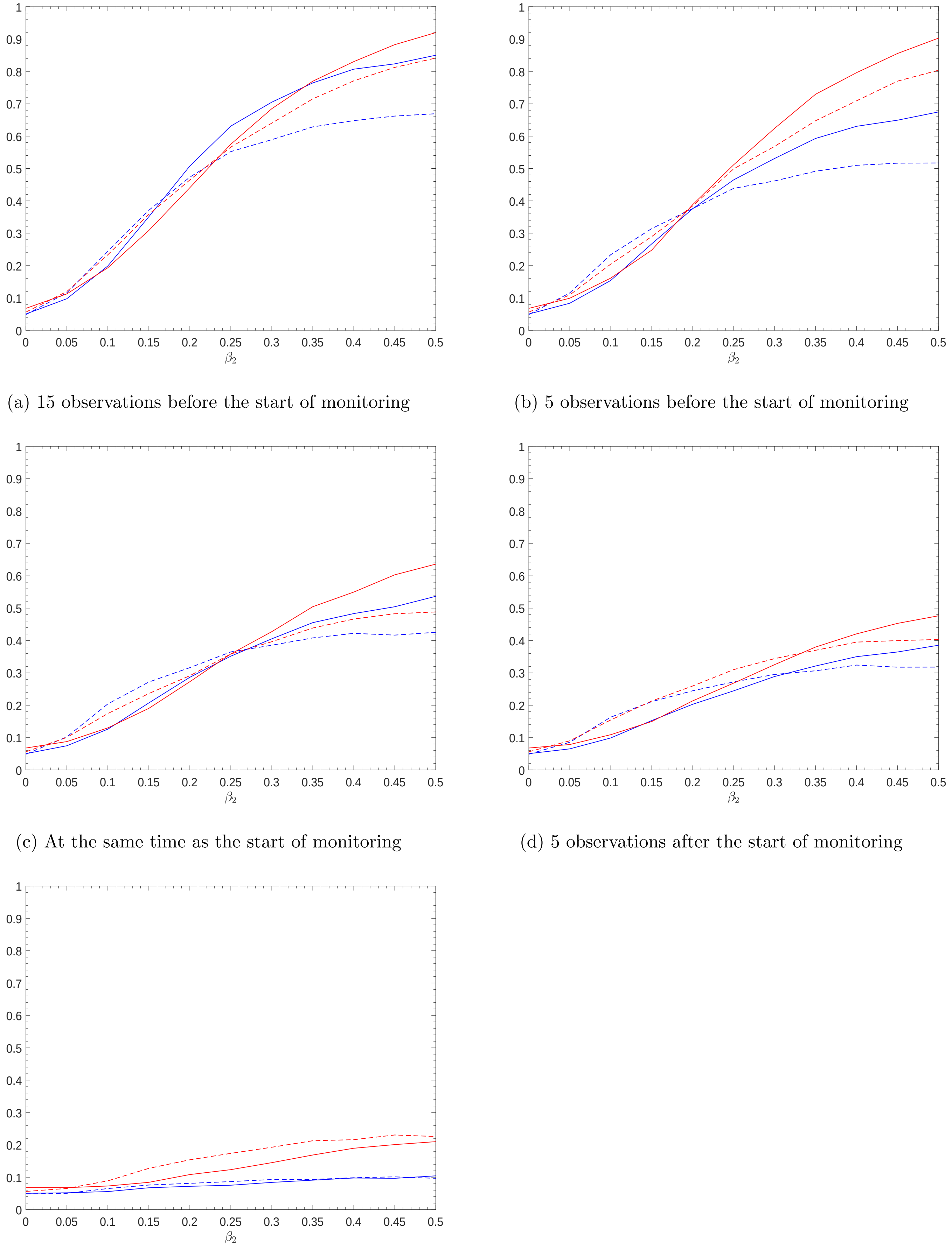
(b) 5 observations before the start of monitoring

(c) At the same time as the start of monitoring

(d) 5 observations after the start of monitoring

(e) 15 observations after the start of monitoring

Figure 3. Predictive regime detection frequency as a function of β_1 for different values of ρ: $T^* + m = 302$, $E = 361$, $m_1 = 30$, $m = 30$; MAX, $\rho = 0.965$, —; MAX, $\rho = 0.995$, - - -; SEQ, $\rho = 0.965$, —; SEQ, $\rho = 0.995$, - - -



(a) 15 observations before the start of monitoring

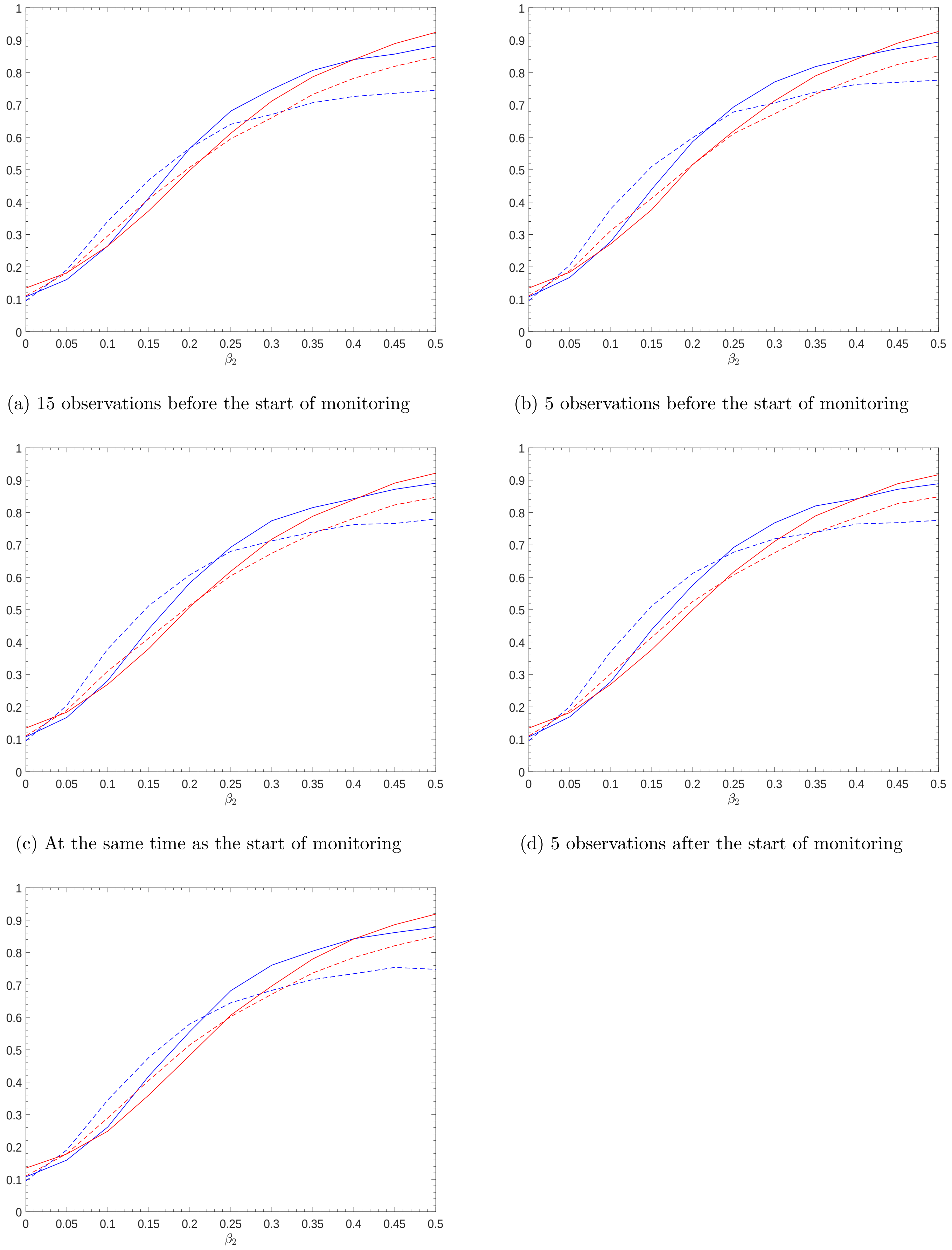
(b) 5 observations before the start of monitoring

(c) At the same time as the start of monitoring

(d) 5 observations after the start of monitoring

(e) 15 observations after the start of monitoring

Figure 4. Detection frequency for a second predictive regime when a predictive regime with $\beta_1 = 0.25$ also exists in the training period, as a function of β_2 for different values of ρ: $T^* + m = 302$, $E = 327$, $m_1 = 15$, $m_2 = 30$, $m = 30$; MAX, $\rho = 0.965$, —; MAX, $\rho = 0.995$, - - -; SEQ, $\rho = 0.965$, —; SEQ, $\rho = 0.995$, - - -



(a) 15 observations before the start of monitoring

(b) 5 observations before the start of monitoring

(c) At the same time as the start of monitoring

(d) 5 observations after the start of monitoring

(e) 15 observations after the start of monitoring

Figure 5. Detection frequency for a second predictive regime when a predictive regime with $\beta_1 = 0.25$ also exists in the training period, as a function of β_2 for different values of ρ: $T^* + m = 302$, $E = 361$, $m_1 = 15$, $m_2 = 30$, $m = 30$; MAX, $\rho = 0.965$, —; MAX, $\rho = 0.995$, - - -; SEQ, $\rho = 0.965$, —; SEQ, $\rho = 0.995$, - - -

frequencies of our procedures to be close to 0.10. It can be seen that each of the curves reported in Figure 2 indeed starts from approximately 0.10. For both the MAX and SEQ procedures, when the predictive regime starts before or at the same time as the start of monitoring (cases (a)-(c)), power rises rapidly with β_1. When the predictive regime starts after the start of monitoring (cases (d)-(e)), a higher proportion of the subsamples used when computing $\tau_{e,m}$ will be data from the period of the DGP when no predictability exists. Furthermore, in these two cases monitoring ends shortly after the predictive regime starts (e.g. for case (e), monitoring ends 11 observations after the predictive regime starts). Therefore, as expected, for both procedures power rises with β_1 at a lower rate than for cases (a)-(c) and ultimately flattens out at a lower value.

Interestingly, these experiments show that the relative finite sample performance of the MAX and SEQ procedures is sensitive to the strength of the predictability (as measured by the magnitude of β_1), the location of the predictability regime relative to the monitoring period, and the persistence of the predictor (as measured by the value of ρ). For case (a), when predictability starts 15 observations before the start of monitoring, and for $\rho = 0.965$, SEQ has more power than MAX, but the difference in power declines as the strength of the predictability increases. Eventually, the power curve for MAX moves above the curve for SEQ (at approximately $\beta_1 = 0.35$). For case (a) with $\rho = 0.995$, the crossing point of the power curves occurs earlier (at approximately $\beta_1 = 0.23$). For case (b) the results have a similar pattern to case (a), although MAX has even more power than SEQ when the predictability is strong compared with case (a). Similar results are found for case (c), although power is noticeably lower for all values of β_1. This is to be expected because in this case the predictability regime starts at the same time as the monitoring, and therefore the initial subsamples used to compute $\tau_{e,m}$ contain very few observations from the predictability regime (by definition, the subsamples used to compute $\tau_{e,m}$ contain more observations from the period when there is no predictability until half way through the monitoring period). For case (d), and $\rho = 0.965$, MAX and SEQ have very similar power when the predictability is weak, although as the predictability strengthens the power curve for MAX moves above the curve for SEQ. The same general pattern exists for $\rho = 0.995$, although the power of both procedures when the predictability is weak is higher than for $\rho = 0.965$. For case (e), MAX has more power than SEQ for all values of β_1, and the difference in power increases as the predictability strengthens.

The results from the second set of experiments are given in Figure 3. As expected, when the monitoring period is extended from $E = 327$ to $E = 361$ the predictive regime detection frequency as a function of β_1 increases for both MAX and SEQ. Indeed the detection frequency and relative finite sample performance of MAX and SEQ are now virtually identical for each of the predictive regime start dates considered here. This reflects the fact that because of the longer monitoring period, each set of sequential $\tau_{e,m}$ statistics now includes a run of statistics computed using subsamples where a high proportion of each subsample is data from when predictability exists in the DGP. When $\beta_1 = 0$ the empirical FPRs of MAX and SEQ both increase to approximately 0.20, again as expected. A further interesting feature of our monitoring procedures can be seen by comparing Figures 2a and 3a relating to the case where the predictive regime starts 15 observations before monitoring begins (of the cases considered, the one where detection power is least dependent on the start date of the predictive regime). Although, as discussed above, the FPR in Figure 3a is roughly double that in Figure 2a for each procedure, very little differences (for a given value of ρ) are seen between the two different cases in terms of the efficacy of MAX and SEQ to detect a predictive regime, except where β_1 is close to zero. Equation (9) shows that, other things being equal, the longer is the length of the training period relative to the monitoring period, the smaller is the theoretical FPR of the procedure. But as these simulation results highlight, a lower FPR from a longer training period does not entail a decrease in the efficacy of the procedures to detect a true predictive regime in the monitoring period.

The results for the third and fourth sets of experiments are given in Figures 4 and 5. We find that, as expected, due to the presence of a predictive regime during the training period, in each of the individual experiments both $\max_{e\in[m+1,T^*]} \tau_{e,m}$ and l_π are increased relative to the case where no predictability is present in the training period, and as a result, the power curves are generally lower in these experiments than the corresponding curves in Figures 2 and 3. For both MAX and SEQ, when $\beta_2 = 0$ and $E = 327$ (consistent with $\alpha = 0.10$), the detection frequency in Figure 4 is approximately 0.05. When $\beta_2 = 0$ and $E = 361$ (consistent with $\alpha = 0.20$), the detection frequency for both procedures in Figure 5 is approximately 0.10. Similarly, it can be seen in Figures 4 and 5 that for $\beta_2 > 0$, the curves are approximately 0.05-0.10 lower than the corresponding curves in Figures 2 and 3. The curves in Figure 4 for $E = 327$ are sensitive to where the second predictive regime is located. However it can be seen in Figure 5 that, as in Figure 3, extending the monitoring period to $E = 361$ reduces the sensitivity of the curves to the exact location of the predictive regime.

4.2 Additional Simulations

The first set of additional simulations studies the detection power of the MAX and SEQ procedures as a function of m_1 (the length of the predictability regime in the DGP), employing the same DGP used in the main experiments and assuming the other parameters are fixed at their original values. The results are graphed in Figure S1 for $E = 327$ and in Figure S2 for $E = 361$. Increases in m_1 from a low value initially lead to an increase in detection power. For larger values of m_1 the power curves flatten out. This occurs because as m_1 increases, eventually the end of the predictability regime in the DGP lies beyond the end of the monitoring period, which in these experiments is assumed to be fixed. When monitoring ends at $E = 327$ the point at which the power curves flatten out occurs earlier as we move from start dates (a) to (e), because the value of m_1 such that the end of the predictive regime lies beyond the end of the monitoring period E gets smaller. For the longer monitoring period $E = 361$ there is very little difference in detection power for the different start dates.

We also carried out an extensive set of robustness checks for MAX and SEQ. The first checks concern the error terms in the DGP. An attractive feature of our monitoring procedure, as Proposition 1 shows, is that for sufficiently large T, in addition to being robust to any degree of contemporaneous correlation of the error terms in the DGP, it is also robust to conditional heteroskedasticity and non-Gaussianity in the errors. To investigate how well these robustness properties hold in finite samples, we repeated the first set of main simulation experiments discussed above using the same DGPs but for a range of error distributions and heteroskedasticity patterns for $\epsilon_{y,t}$ in (1): (i) $t(10)$ error terms; (ii) $t(5)$ error terms; (iii) normally distributed GARCH(1,1) error terms with conditional variance $\sigma^2_{y,t} = \alpha_0 + \alpha_1 \epsilon^2_{y,t-1} + \beta_1 \sigma^2_{y,t-1}$ where $\alpha_0 = 0.10$, $\alpha_1 = 0.10$ and $\beta_1 = 0.80$, and (iv) $t(5)$ GARCH(1,1) error terms with the same GARCH parameters. Although not formally allowed under the conditions of Proposition 1, we also considered: (v) $t(5)$ error terms with an unconditional volatility shift from $\sigma_y = 1$ to $\sigma_y = 2$ halfway through the monitoring period (at $t = T^* + m + \lfloor (E - T^* - m)/2 \rfloor + 1$). Reassuringly, the results, which are graphed in Figures S3-S7, are very similar to the first set of main simulation results reported in Figure 2. As discussed in Remark 3, the AR(1) specification for the predictor in (1)-(3) is not critical for our analysis, and for large T, both the MAX and SEQ procedures remain valid for higher order autoregressive predictors. To investigate this issue in finite samples we report the results from repeating the first set of main simulation experiments given in Figure 2, but using an AR(2) predictor rather than an AR(1); that is replacing the AR(1) process in (3) by $s_{x,t} = \rho_1 s_{x,t-1} + \rho_2 s_{x,t-2} + \epsilon_{x,t}$, $t = 1, ..., T$, setting $\rho_1 = 0.595$, and allowing $\rho_2 = \{0.30, 0.40\}$. The results are given in Figure S8 and again they are very similar to the main simulation results reported in Figure 2.

As a final robustness check, we investigated the detection power of MAX and SEQ when the regime change in (1)-(3) is gradual rather than discrete. Specifically, we used the DGP for the first set of main simulation experiments but redefined the dummy variable $d_t(e_1, m_1)$ to be the exponential function $d_t(e_1, m_1) := exp(-\gamma(t - s)^2)$, which allows for smooth regime change centered around s, where γ controls the speed of the change. We set $s = e_1 - 0.50m_1 + 1$ and $\gamma = 0.01$ so that the main part of the regime change for cases (a)-(e) starts at approximately the same point as in Figure 2 and lasts for approximately 30 observations. The results are given in Figure S9 and show that as β_1 increases, both MAX and SEQ have good detection power for this form of regime change. Generally the rate of increase in power with increases in β_1 is slower than in Figure 2 and the curves are slower to flatten out, which occurs because increases in β_1 are effectively being weighted by a factor less than one for most of the predictability regime; hence, the β_1 that maximises power (assuming the other parameters in the DGP are fixed) is higher than for the results in Figure 2.

5 Empirical Application

5.1 Data and Preliminary Analysis

The dataset used for the empirical application of our monitoring procedure consists of monthly observations on the equity premium for the S&P Composite index calculated using CRSP's month-end values and on 20 different predictors for the period 1974:12-2015:12 ($T = 493$). We define the equity premium as in Goyal and Welch (2008) and Neely *et al.* (2014) as the log return on the value-weighted CRSP stock market index minus the log return on the risk-free Treasury bill: $y_t = log(1 + R_{m,t}) - log(1 + R_{f,t})$ where $R_{m,t}$ is the CRSP return and $R_{f,t}$ is the Treasury bill return. Ten of the predictors are traditional macroeconomic and financial variables (MFVs) and ten are binary technical analysis indicators (TAIs) also used by Neely *et al.* (2014) in their analysis of equity premium predictability. The traditional MFVs are in log form (as in Goyal and Welch, 2008; Neely *et al.*, 2014) and each of the predictors is lagged one period. We consider the log dividend yield (dy_{t-1}), the log dividend-price ratio (dp_{t-1}), log earnings-price ratio (ep_{t-1}), book-to-market ratio (bm_{t-1}), short term yield (st_{t-1}), long-term yield (lt_{t-1}), long-term - short-term yield spread ($sp_{t-1} = lt_{t-1} - st_{t-1}$), BAA-AAA corporate bond yield spread (dsp_{t-1}), net equity expansion ($ntis_{t-1}$), and inflation (inf_{t-1}). The TAIs used are four moving average indicators (MAIs), two momentum indicators (MOIs), and four on-balance volume (OBV) indicators. The four moving-average rule indicators ($MAI_{s,l,t}$) are defined such that $MAI_{s,l,t} := 1$ if $MA_{s,t} \geq MA_{l,t}$, indicating a buy signal, and are defined to be zero otherwise, where $MA_{j,t} := (1/j) \sum_{i=0}^{j-1} P_{t-i}$ for $j = \{s, l\}$ and $s = \{1, 2\}$, $l = \{9, 12\}$ and where P_t is the level of the S&P Composite index. The two l-period momentum rule indicators ($MOI_{l,t}$) are defined such that $MOI_{l,t} := 1$ if $P_t \geq P_{t-l}$, indicating a buy signal, and are defined to be zero otherwise, where $l = \{9, 12\}$. The four on-balance volume rule indicators ($OBV_{s,l,t}$) are defined such that $OBV_{s,l,t} := 1$ if $MA^{OBV}_{s,t} \geq MA^{OBV}_{l,t}$, indicating a buy signal, and are defined to be zero otherwise, where $MA^{OBV}_{j,t} := (1/j) \sum_{i=0}^{j-1} obv_{t-i}$ for $j = \{s, l\}$ and $s = \{1, 2\}$, $l = \{9, 12\}$, and, $obv_t := \sum_{k=1}^{t} VOL_k D_k$, where VOL_k is trading volume for the S&P Composite index in period k and D_k is a binary variable such that $D_t := 1$ if $P_t \geq P_{t-1}$ and $D_t := -1$ otherwise.

The data used to construct the equity premium and the predictors are taken from the updated monthly data set on Amit Goyal's website (`www.hec.unil.ch/agoyal/`) which is an extended version of the data set used by Welch and Goyal (2008). A full list of the predictors is given in Table S1 of the supplementary appendix.

We begin with a preliminary analysis using some popular orthodox methods for detecting predictability. Table S2 in the supplementary appendix reports, for each predictor variable considered, the estimated slope parameter ($\hat{\beta}$), a right-tailed Newey-West t-test of significance (t_{NW}) and the standard and adjusted R^2 values for orthodox bivariate regression models applied to the full sample of data using OLS for parameter estimation. For both the MFVs and the

TAIs, consistent with many of the previous empirical studies discussed in section 1 very little evidence of predictability is provided by the t_{NW} tests run at conventional significance levels and in all cases the R^2 values are under 1%. It is important to recognize that although popular in studies of equity premium predictability, orthodox t-tests (including t_{NW}) can be misleading in this case because of the highly persistent lagged regressors used (see again the discussion in section 1), therefore also reported in Table S2 is the IV_{comb} test of Breitung and Demetrescu (2015). This statistic has a standard normal asymptotic null distribution, such that the test is valid, irrespective of the persistence of the predictor and any heteroskedasticity present in the errors. As discussed in Remark 4 of Breitung and Demetrescu (2015,p.364), the IV_{comb} test can only be validly implemented as a two-tailed test. For the MVFs there is no statistically significant evidence of predictability from IV_{comb} at conventional significance levels, and only a single rejection at the 0.10 significance level for the TAIs.[10]

Recall that in outlining our monitoring procedure in section 3 we assumed in generating the empirical critical value, cv_π, that there was no predictability over the training periods. To assess how this assumption sits with our data sets we apply the same methods used for obtaining the full sample results in Table S2 to the training periods employed in the monitoring application below. Although we present the results for all of the methods used in Table 2, to assess the presence of predictability in these training periods we focus on the IV_{comb} test. For the monitoring application below, our initial choice of training periods is 12/74-10/98 (for $m = 15$), 12/74-07/97 (for $m = 30$), and 12/74-01/95 (for $m = 60$). These are the implied training periods given by $T^* = 302 - m$, where observation $t = 302$ is the date at which monitoring starts in the application below, 01/00. If there is statistically significant evidence of predictability for an initial choice of training period, but this is thought to be due to a period of predictability towards the end of that training period, then we recommend ending the training period at an earlier date so as to reduce the likelihood it contains predictability. Thus, the final training periods employed when monitoring could finish earlier than the initial choice of training period; see the discussion in subsection 3.4.[11]

Our preliminary analysis of the data over the implied training periods reveals that for the two interest rate series st_{t-1} and lt_{t-1}, and for the bond yield spread dsp_{t-1}, there is statistically significant evidence of predictability at conventional significance levels from IV_{comb} for one or more values of m. Furthermore, the rejections obtained do not appear to be driven by predictability at the end of these implied samples. Therefore, in the monitoring application below we continue

[10] Financial theory suggests negative predictive power for st_{t-1}, lt_{t-1}, $ntis_{t-1}$ and inf_{t-1}. We therefore multiply each of these predictors by -1 so that a right-sided test (excepting the IV_{comb} test which, as discussed above, is implemented as a two-tailed test) is appropriate for detecting predictability. See footnote 5 for further details.

[11] If predictability is present during the training period, as the simulations in section 4 demonstrate, our procedure can still be useful for detecting positive predictability over the monitoring period. Note that if negative predictability exists over the training period and a predictability regime change is detected using the upper-tailed version of our procedure, we cannot conclude that the change is to a period of positive predictability without further analysis, because it could be due to a change to a period of no, or less negative, predictability.

to use the implied training periods for these three predictors despite the rejections from IV_{comb}. Statistically significant evidence of predictability from IV_{comb} is also obtained for $ntis_{t-1}$, for all values of m. In this case, we find that predictability is concentrated in the data from 01/92 through to the end of the training periods. Hence, for this predictor and for all values of m, we end the relevant training periods at 12/91 in the monitoring application below. For all of the other MFV and TAI predictors no statistically significant evidence of predictability is found from IV_{comb} using the implied training periods. The full set of results from the preliminary analysis of the data over the training periods (using the adjusted training period for $ntis_{t-1}$) are given in Tables S3 and S4 in the supplementary appendix for the MFVs and TAIs respectively.

5.2 Monitoring Results

We assume that a practitioner applies our MAX and SEQ procedures to monitor for the emergence of predictive regimes from 01/00 (so in all cases $T^* + m = 302$). Results are presented assuming that monitoring continues through to the final data observation, 12/15. In real-world applications it is not envisaged that our procedures would be used for continuous monitoring over anything like such a long period, but it is helpful to present the results through to 12/15 to illustrate the relationship between the length of the monitoring period and the FPR. Results are computed for $m = \{15, 30, 60\}$. For the SEQ procedure we have computed results for both 0.10 and 0.05 level estimated critical values, i.e. cv_π for $\pi = \{0.10, 0.05\}$, but we concentrate here on the results for $\pi = 0.10$. The results for $\pi = 0.05$ are given in Table S5 and Table S6 of the supplementary appendix.

Table 1 reports the number of predictive regimes detected by MAX and SEQ (with $\pi = 0.10$) respectively. For each predictor where one or more predictive regimes are detected, Table 2 reports the date at which the first regime is detected and the associated empirical FPR for both MAX and SEQ (using $cv_{0.10}$). Notice that the TAI predictors are 0-1 dummy variables that will often take the same value for several consecutive observations, and consequently the subsample $\tau_{e,m}$ values can be undefined when the TAI does not change over the subsample. If $\tau_{e,m}$ is undefined during the monitoring period it simply means that at the relevant observation when this occurs the test statistic is uninformative about the presence of predictability, but the $\tau_{e,m}$ values that *are* defined can still be used for monitoring. However, a large number of undefined test statistics in the training period could have a detrimental impact on the finite sample performance of the procedure. For completeness, the results for $m = \{15, 30\}$ are reported in these tables, although for some of the TAIs undefined test statistics occur quite frequently over the training period with these values of m. In practice, we recommend using $m \geq 60$ when using our procedure with these particular TAIs to minimize the number of undefined test statistics over the training period. Alternatively, for a given value of m, reducing the value of l when constructing the TAIs will result in fewer undefined test statistics. In the application here we report results for $l = \{9, 12\}$ to be consistent with the regression-based analysis of TAIs in Neely *et al.* (2014), even though for some of the MOIs and OBV indicators with $m = 60$ and these values of l, $\tau_{e,m}$ is occasionally undefined over the training and/or monitoring period. For the MAIs with $l = \{9, 12\}$ and $m = 60$ there are no undefined test statistics.

Table 1. Number of predictive regimes detected by MAX and SEQ

	MAX			SEQ, $\pi = 0.10$		
	$m = 15$	$m = 30$	$m = 60$	$m = 15$	$m = 30$	$m = 60$
			MFVs			
dy_{t-1}	0	2	1	1	0	2
dp_{t-1}	0	1	1	1	0	3
ep_{t-1}	1	2	1	2	3	1
bm_{t-1}	1	0	2	1	0	1
st_{t-1}	0	1	1	0	0	0
lt_{t-1}	0	3	1	3	1	2
sp_{t-1}	0	0	0	0	0	0
dsp_{t-1}	1	2	1	1	0	0
$ntis_{t-1}$	0	1	0	0	0	0
inf_{t-1}	0	0	0	2	0	0
			TAIs			
$MAI_{1,9,t-1}$	0	3	3	0	1	2
$MAI_{1,12,t-1}$	2	1	3	0	2	3
$MAI_{2,9,t-1}$	1	3	2	0	1	2
$MAI_{2,12,t-1}$	1	1	3	0	2	3
$MOI_{9,t-1}$	0	0	3	0	3	2
$MOI_{12,t-1}$	0	3	3	0	2	2
$OBV_{1,9,t-1}$	0	2	0	0	1	1
$OBV_{1,12,t-1}$	1	0	2	1	2	1
$OBV_{2,9,t-1}$	1	0	1	0	0	2
$OBV_{2,12,t-1}$	0	0	3	1	1	3

It can be seen from Table 1 that, in total, employing the three sub-sample sizes $m = \{15, 30, 60\}$ leads to one or more predictive regimes being detected by MAX for eight of the ten MFVs. MAX finds no evidence of predictability for sp_{t-1} and inf_{t-1}. Notice that the total number of MFVs found to have predictive power is lower for $m = 15$ than for the larger values of m considered. In total, one or more predictive regimes are detected for all ten of the TAIs considered and the number of TAIs found to have predictive power increases with m; from five for $m = 15$, to six for $m = 30$, and nine for $m = 60$.

Consider now the results from using the SEQ procedure, also given in Table 1. In total, employing the three sub-sample sizes $m = \{15, 30, 60\}$, one or more predictive regimes are detected by SEQ for seven of the ten MFVs, and for all of the TAIs. Notice that in contrast to MAX, the total number of MFVs found to have predictive power is largest for $m = 15$:

Table 2. First month where a predictive regime is detected by MAX and SEQ

	MAX						$SEQ, \pi = 0.10$					
	$m = 15$		$m = 30$		$m = 60$		$m = 15$		$m = 30$		$m = 60$	
	MAX	FPR_{MAX}	MAX	FPR_{MAX}	MAX	FPR_{MAX}	SEQ	FPR_{SEQ}	SEQ	FPR_{SEQ}	SEQ	FPR_{SEQ}
MFVs												
dy_{t-1}	N/A	N/A	02/01	0.055	02/14	0.483	09/07	0.255	N/A	N/A	02/02	0.125
dp_{t-1}	N/A	N/A	02/01	0.055	02/14	0.483	05/15	0.405	N/A	N/A	01/02	0.121
ep_{t-1}	07/11	0.338	01/08	0.286	01/09	0.375	09/03	0.142	01/04	0.168	12/04	0.248
bm_{t-1}	07/00	0.025	N/A	N/A	07/01	0.095	10/00	0.035	N/A	N/A	02/02	0.125
st_{t-1}	N/A	N/A	03/11	0.358	10/12	0.458	N/A	N/A	N/A	N/A	N/A	N/A
lt_{t-1}	N/A	N/A	04/03	0.142	03/05	0.257	10/03	0.145	08/04	0.188	08/05	0.272
sp_{t-1}	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
dsp_{t-1}	07/12	0.357	08/11	0.366	02/14	0.483	05/12	0.354	N/A	N/A	N/A	N/A
$ntis_{t-1}$	N/A	N/A	08/11	0.444	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
inf_{t-1}	N/A	N/A	N/A	N/A	N/A	N/A	06/04	0.166	N/A	N/A	N/A	N/A
TAIs												
$MAI_{1,9,t-1}$	N/A	N/A	12/07	0.284	10/08	0.368	N/A	N/A	11/09	0.330	06/04	0.229
$MAI_{1,12,t-1}$	11/00	0.039	01/08	0.286	09/02	0.153	N/A	N/A	01/04	0.168	09/02	0.153
$MAI_{2,9,t-1}$	01/08	0.263	11/08	0.307	10/08	0.368	N/A	N/A	07/10	0.344	10/05	0.278
$MAI_{2,12,t-1}$	01/08	0.263	01/08	0.286	09/01	0.103	N/A	N/A	02/04	0.171	04/02	0.133
$MOI_{9,t-1}$	N/A	N/A	N/A	N/A	10/00	0.052	N/A	N/A	10/03	0.160	02/02	0.125
$MOI_{12,t-1}$	N/A	N/A	08/03	0.154	12/00	0.062	N/A	N/A	06/04	0.182	12/01	0.117
$OBV_{1,9,t-1}$	N/A	N/A	11/08	0.307	N/A	N/A	N/A	N/A	06/09	0.320	05/10	0.407
$OBV_{1,12,t-1}$	01/08	0.263	N/A	N/A	02/09	0.377	11/09	0.304	06/04	0.182	12/09	0.397
$OBV_{2,9,t-1}$	01/08	0.263	N/A	N/A	02/09	0.377	N/A	N/A	N/A	N/A	06/10	0.409
$OBV_{2,12,t-1}$	N/A	N/A	N/A	N/A	11/08	0.370	11/09	0.304	03/10	0.337	10/01	0.108

predictive regimes are detected for seven MFVs when $m = 15$, two when $m = 30$, and five when $m = 60$. The total number of TAIs found to have predictive power increases with m, from two for $m = 15$, to nine for $m = 30$, and ten for $m = 60$. Our results from both MAX and SEQ are, in general, consistent with the findings in Neely *et al.* (2014), that stronger evidence of predictability is found for the TAIs than for the MFVs.

It can be seen in Table 2 that for many of the MFV and TAI predictors, a predictive regime is first detected around the time of the dot-com bubble/crash in the late-1990s/early 2000s, or the global financial crisis in 2008-2009. Table 2 also shows that, as might be expected, for some of the predictors our procedures detect a predictability regime around the same date, and in some cases, in the same month. Consider for example the results using MAX with $m = 30$. For both dy_{t-1} and dp_{t-1}, predictability is first detected in 02/01. For dsp_{t-1} and $ntis_{t-1}$, in both cases predictability is first detected in 08/11.

The dating procedures discussed in sub-section 3.3 also provide useful information on the location of the regimes. As an example, Figure 6 graphs $\tau_{e,m}$ along with the weak set of dates obtained using MAX with $m = 30$ for the dividend-price ratio dp_{t-1} as a predictor (note that the strong set of dates is empty in this case). Figures 7 and 8 graph the MAX results with $m = 30$ for the short-term and long-term interest rates, st_{t-1} and lt_{t-1}. A selection of graphical results for the other predictors for which at least one predictive regime is signalled for either the MAX or SEQ procedures are provided in the supplementary appendix in Figures S10-S13. For presentational purposes, in these graphs we do not display $\tau_{e,m}$ over the entire training period and instead start the horizontal axis five years before the end of each training period. Also indicated on these graphs are the end of the training period T^*, the date when monitoring starts $T^* + m$, the largest $\tau_{e,m}$ in the training period ($\max_{e \in [m+1,T^*]} \tau_{e,m}$), the date of the first significant rejection for the i-th predictive regime j_i (for the MAX procedure, this is the date at which the i-th predictive regime is detected), and the FPR, based on (9), as a function of E.

Figure 6 shows that for dp_{t-1}, the MAX procedure with $m = 30$ detects a single predictive regime in 02/01 and the weak set of dates covers the period 09/98-03/01. Thus our results suggest that dp_{t-1} had predictive power for equity returns during the latter years of the dot-com bubble period. Notice that the weak set of dates starts before the monitoring period, which can happen for early rejections because the rejection itself is indexed on the end date of the sub-sample window. For st_{t-1}, the MAX procedure with $m = 30$ detects one predictive regime and Figure 7 shows that the weak set of dates cover the period 10/08-03/11. However, for lt_{t-1} the MAX procedure with $m = 30$ detects three predictive regimes. Figure 8 shows that in this case, the weak set of dates covers the periods 11/00-11/04, 03/11-08/13, 05/11-11/13. Therefore, the weak set of dates associated with the second and third regimes overlap - suggesting a single period of predictability that begins in 03/11 and ends in 11/13. Figure 8 shows that the first regime detected by MAX for lt_{t-1} in 04/03 follows a gradual increase in $\tau_{e,m}$ that began after the dot-com crash and continued through to late-2004. Over this period U.S. interest rates gradually

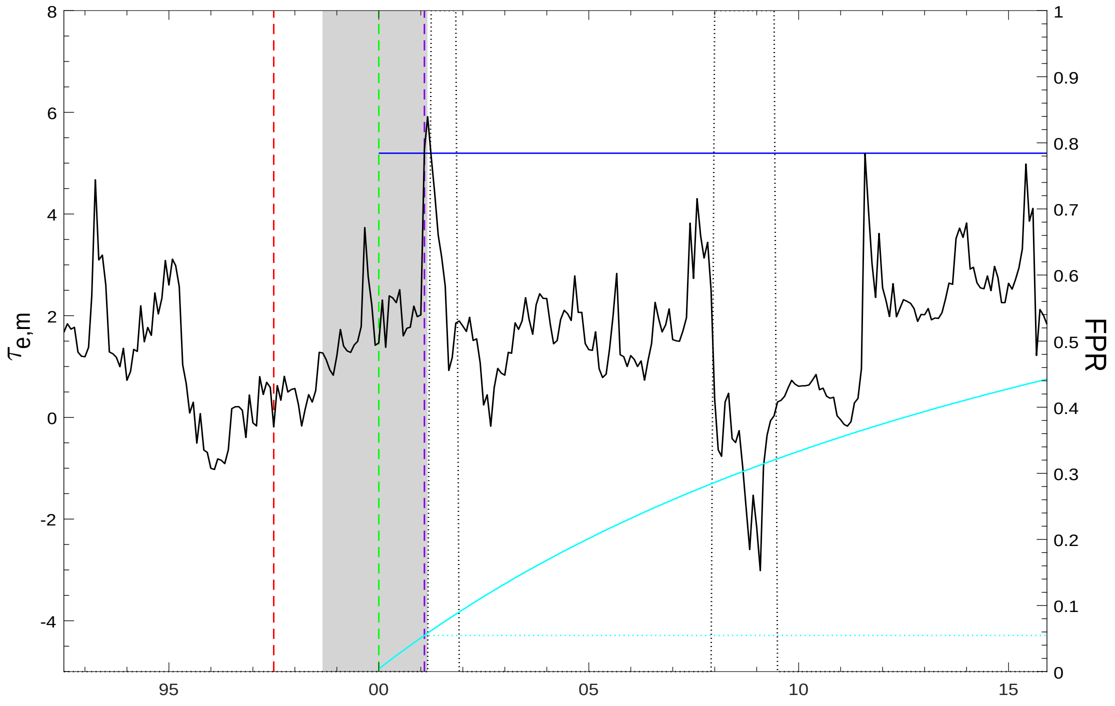


Figure 6. dp_{t-1}, MAX procedure, $m = 30$: —($\tau_{e,m}$), —($\max_{e\in[m+1,T^*]} \tau_{e,m}$), - - -($T^*$), - - -($T^* + m$), - - -(first rejection), ▨ (weak set of dates), —(false positive rate), ⋯⋯(NBER indicator)

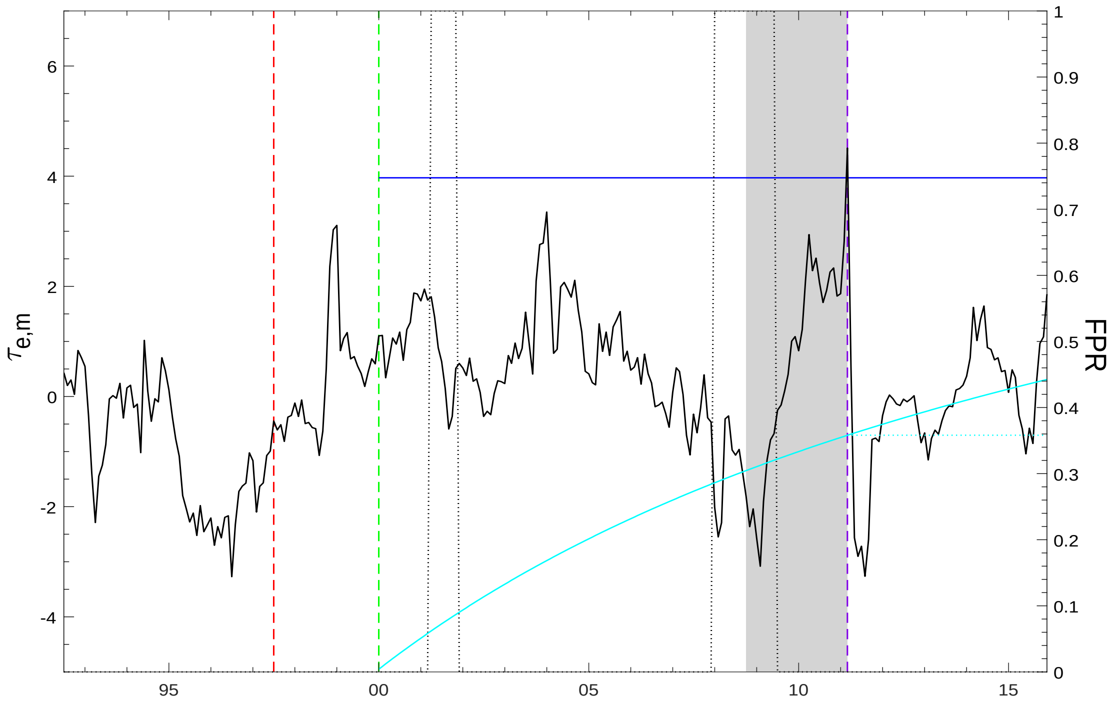


Figure 7. st_{t-1}, MAX procedure, $m = 30$: —($\tau_{e,m}$), —($\max_{e\in[m+1,T^*]} \tau_{e,m}$), - - -($T^*$), - - -($T^* + m$), - - -(first rejection), ▨ (weak set of dates), —(false positive rate), ⋯⋯(NBER indicator)

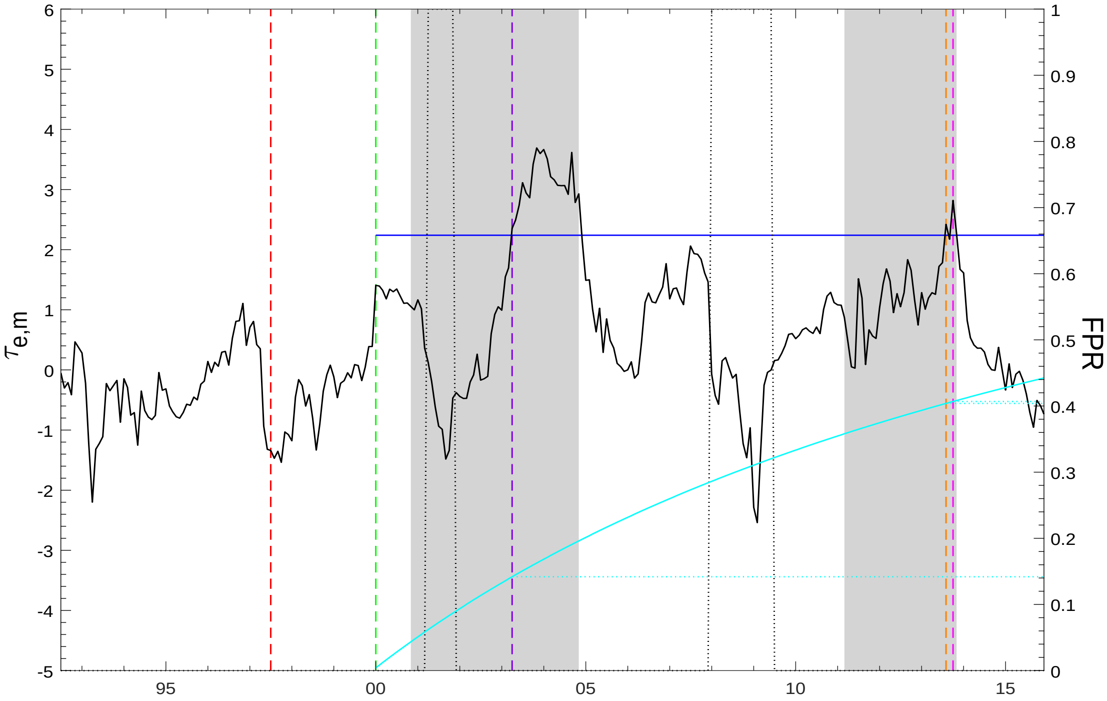


Figure 8. lt_{t-1}, MAX procedure, $m = 30$: —($\tau_{e,m}$), —($\max_{e\in[m+1,T^*]}\tau_{e,m}$), - - -($T^*$), - - -($T^* + m$), - - -(first rejection), - - -(second rejection), - - -(third rejection), ▨ (weak set of dates), —(false positive rate), ⋯⋯(NBER indicator)

fell and equity markets recovered after the dot-com crash and 2001 recession. Our results suggest that the long-term interest rate had predictive power over this period but the short-term interest rate did not. The second and third regimes for lt_{t-1} are shorter in duration than the first and are largely driven by a rapid and short-lived increase in $\tau_{e,m}$ during 2013.

Neely *et al.* (2014) investigate differences in predictability between macroeconomic recession and expansion periods by computing separate R^2 statistics for predictive regression models using the NBER indicator of recessions and expansions to partition the relevant data. They find that for both the MFVs and TAIs predictability is substantially higher over recessions than over expansions. In the light of these findings it is interesting to compare the subsample $\tau_{e,m}$ values over the monitoring period with the NBER indicator to see if our procedure finds a similar pattern of support for predictability over the business cycle. Hence, the NBER indicator is also plotted in Figures 6-8. There are two US recessions over the monitoring period 01/00-12/15: one short recession in early 2001 (March 2001-November 2001), and one major recession associated with the global financial crisis (December 2007-June 2009). Figure 6 shows that for dp_{t-1}, predictability peaks at the start of the 2001 recession but declines during the course of the recession; for st_{t-1} and lt_{t-1} the predictive regimes detected do not appear to be correlated with the business cycle. As shown in the supplementary appendix, for the other predictors, whilst there is some evidence suggesting that, consistent with the findings in Neely *et al.* (2014), predictability is stronger during recessions than during expansions, it is not a pattern obtained for all of the predictors.[12] It is interesting to relate our results to recent research by Farmer *et al.* (2019), who also focus on detecting short pockets of in-sample predictability in U.S. equity returns. While the sample sizes and the number of predictors they analyse differ from ours, there are some similarities between their results and ours. For example, for the dividend yield, Farmer *et al.* (2019) find evidence of pockets of predictability in the early-2000s and the early/mid-2010s; and for the Treasury bill rate in the late-2000s and the early/mid-2010s. These dates are similar to the predictive regime dates obtained for these predictors using our MAX procedure.

The predictive regimes in Figures 6-8 often end quite shortly after each regime is first detected (e.g. in Figure 7, the weak set of dates ends immediately after the regime is detected). Indeed, this general pattern was observed for all of the MFVs and for the majority of the TAIs. Hence, the strong set of dates for most of the predictors is empty. This suggests that although investors using our procedure in real time would have been able to detect predictability in these cases, there may have been very little time after the point of detection to exploit the predictability before it no longer existed. To investigate this point using traditional forecasting methods, for each MFV predictor where one or more predictive regimes are detected by MAX and/or SEQ with $m = 30$ we computed out-of-sample forecasts exploiting the information from the monitoring procedures. Specifically, for each of these predictors we move forward through the monitoring period one month at a time computing MAX and SEQ at each month along with one step-ahead forecasts. To compute the forecasts we use a fixed mean benchmark model estimated using an expanding sample of data that starts at the first observation, until the relevant monitoring procedure detects a first predictive regime. When this occurs we use the relevant regression model to compute the forecast for the next month, estimated using an expanding sample of data that starts at the weak start date for the relevant predictive regime. When the first predictive regime ends, we stop forecasting. We compared the forecasts computed in this way with the forecasts obtained using the fixed mean benchmark model for the whole forecasting period. The mean squared forecast error (MSFE) for each procedure, along with the Diebold and Mariano (1995) test of equal forecasting accuracy (employing the Harvey *et al.*, 1997, bias-correction and Student's t-critical values), and the out-of-sample R^2 value for the procedure are reported in Table S7 in the supplementary appendix. As expected, because the predictive regimes end so quickly after they are discovered, for the majority of predictors there is very little difference between the MSFE obtained exploiting our MAX and SEQ procedures in this way and the MSFE for the benchmark model. In some cases the MSFE using our test in this way is lower than the benchmark model, but the differences are not statistically significant. Paye and Timmermann (2006) and Timmermann (2008) argue that if predictability reflects market inefficiencies then it

[12] We note that Neely *et al.* (2014) study a longer sample of the data than the sample used here that ends earlier (12/50-12/11) and their empirical work is fundamentally different to ours, being an *ex post* analysis of predictability (in-sample and out-of-sample) rather than a real-time monitoring application.

is only ever likely to be a short-lived phenomenon because when it exists, investors will quickly allocate capital to exploit its presence. Our finding of short pockets of predictability that end quickly after being detected is entirely consistent with this view.

6 Conclusions

We have developed new real-time monitoring procedures for detecting the emergence of predictive regimes. Our detection procedures are based on the sequential application of standard heteroskedasticity-robust (predictive) regression t-statistics for predictability to end-of-sample data. We have suggested two possible detection rules, both of which are designed to be robust to both the degree of persistence and endogeneity of the regressors in the predictive regression and are such that their false positive rates can be controlled, for a given monitoring period length, by using information obtained from data in a training period. We have applied our proposed monitoring procedures to investigate for the presence of regime changes in the predictability of the U.S. equity premium at the one-month horizon by traditional macroeconomic and financial variables, and by binary technical analysis indicators. Our results suggest that the one-month ahead equity premium has displayed episodes of temporary predictability and that these episodes could have been detected in real-time by practitioners using our proposed methodology.

References

Andrews, D.W.K. (2003). End-of-sample instability tests, Econometrica, 71, 1661-1694.

Andrews, D.W.K. and Kim, J-S. (2006). Tests for cointegration breakdown over a short time period, Journal of Business & Economic Statistics, 24, 379-393.

Ang, A. and Bekaert, G. (2007). Stock return predictability: is it there? Review of Financial Studies, 20, 651-707.

Astill, S., Harvey, D.I., Leybourne, S.J., Sollis, R. and Taylor, A.M.R. (2018). Real-time monitoring for explosive financial bubbles, Journal of Time Series Analysis, 39, 863-891.

Bai, J. and Perron, P. (1998). Estimating and testing linear models with multiple structural changes, Econometrica, 66, 47-78.

Bai, J. and Perron, P. (2003). Computation and analysis of multiple structural change models, Journal of Applied Econometrics, 18, 1-22.

Boudoukh, J.R., Michaely, M., Richardson, P. and Roberts, M.R. (2007). On the importance of measuring payout yield: Implications for empirical asset pricing, Journal of Finance, 62, 877-915.

Breitung, J. and Demetrescu M. (2015). Instrumental variable and variable addition based inference in predictive regressions, Journal of Econometrics, 187, 358-375.

Brock, W., Lakonishok, J. and LeBaron, B. (1992). Simple technical trading rules and the stochastic properties of stock returns, Journal of Finance, 47, 1731-1764.

Campbell, J.Y. (1987). Stock returns and the term structure, Journal of Financial Economics 18, 373-400.

Campbell, J.Y. and Shiller, R.J. (1988a). Stock prices, earnings and expected dividends, Journal of Finance, 43, 3, 661-676.

Campbell, J.Y. and Shiller, R.J. (1988b). The dividend-price ratio and expectations of future dividends and discount factors, Review of Financial Studies, 1, 195-228.

Campbell, J.Y. and Thompson, S.B. (2008). Predicting excess stock returns out of sample: can anything beat the historical average? Review of Financial Studies, 21, 1509-1531.

Cochrane, J.H. (2008). The dog that did not bark: a defense of return predictability, Review of Financial Studies, 21, 1533-1575.

Dangl, T. and Halling, M. (2012). Predictive regressions with time-varying coefficients, Journal of Financial Economics, 106, 157–181.

Diebold, F.X. and Mariano, R.S. (1995). Comparing predictive accuracy, Journal of Business and Economic Statistics, 13, 253–263.

Fama, E.F. (1981). Stock returns, real activity, inflation, and money, American Economic Review, 71, 545-565.

Fama, E.F. (1990). Stock returns, expected returns, and real activity, Journal of Finance, 45, 1089-1108.

Fama, E.F. and French, K.R. (1988). Dividend yields and expected stock returns, Journal of Financial Economics, 22, 3–25.

Fama, E.F. and French, K.R. (1989). Business conditions and expected returns on stocks and bonds, Journal of Financial Economics, 25, 23–49.

Farmer, L.E., Schmidt, L. and Timmermann, A. (2019). Pockets of predictability, Discussion Paper, available at SSRN: `https://ssrn.com/abstract=3152386`.

Ferreira, H. and Scotto, M. (2002). On the asymptotic location of high values of a stationary sequence, Statistics and Probability Letters, 60, 475–482.

Goyal, A. and Welch, I. (2003). Predicting the equity premium with dividend ratios, Management Science, 49, 639-654.

Harvey, D.I., Leybourne, S.J. and Newbold, P. (1997). Testing the equality of prediction mean squared errors, International Journal of Forecasting, 13, 281–291.

Henkel, S.J., Martin, J.S. Martin and Nardari, F. (2011). Time-varying short-horizon predictability. Journal of Financial Economics, 99, 560–580.

Inoue, A. and Rossi, B. (2005). Recursive predictability tests for real-time data, Journal of Business and Economic Statistics, 23, 336-345.

Keim, D.B. and Stambaugh, R.F. (1986). Predicting returns in the stock and bond markets, Journal of Financial Economics, 17, 357-390.

Kostakis, A., T. Magdalinos and M.P. Stamatogiannis (2015). Robust econometric inference for stock return predictability. Review of Financial Studies, 28, 1506–1553.

Lettau, M. and Ludvigsson, S. (2001). Consumption, aggregate wealth, and expected stock returns. Journal of Finance, 56, 815-850.

Neely, C.J., Rapach, D.E., Tu, J. and Zhou, G. (2014). Forecasting the equity risk premium: the role of technical indicators, Management Science, 60, 1772-1791.

Nelson C.R. and Kim M.J. (1993). Predictable stock returns: the role of small sample bias, Journal of Finance, 48, 641-661.

Paye, B.S. and Timmermann, A. (2006). Instability of return prediction models, Journal of Empirical Finance, 13, 274-315.

Stambaugh, R.F. (1999). Predictive regressions, Journal of Financial Economics, 54, 375-421.

Timmermann, A. (2008). Elusive return predictability, International Journal of Forecasting, 24, 1-18.

Welch, I. and Goyal, A. (2008). A Comprehensive look at the empirical performance of equity premium prediction, Review of Financial Studies, 21, 1455-1508.

White, H. (1982). Maximum likelihood estimation of misspecified models, Econometrica, 50, 1-25.

Supplementary Appendix

Proof of Proposition 1

The stated result in Proposition 1 follows using an application of Theorem 2.1 of Ferreira and Scotto (2002,p.478) provided $\{\tau_{e,m}\}$ forms a strictly stationary sequence of mixing random variables; the precise mixing conditions required are detailed on page 476 of Ferreira and Scotto (2002). The result in Theorem 2.1 of Ferreira and Scotto (2002) gives the limiting probability that the rth ($r \geq 1$) largest value in any one of two disjoint subintervals of the sample data exceeds the sth ($s \geq 1$) largest value in the other subinterval. We therefore now establish that these two conditions hold on $\{\tau_{e,m}\}$ under the conditions of Proposition 1.

(i) $\{\tau_{e,m}\}$ *is a strictly stationary sequence.* This arises from the strict stationarity of ϵ_t and because we can write $x_{t-1} - \bar{x}_{-1}$ in (6) and (7), for $t = e - m + 1, ..., t = e$, as

$$\begin{aligned} x_{t-1} - \bar{x}_{-1} &= x_{t-1} - m^{-1} \sum_{t=e-m+1}^{e} x_{t-1} \\ &= (x_{t-1} - x_{e-m-1}) - m^{-1} \sum_{t=e-m+1}^{e} (x_{t-1} - x_{e-m-1}) \\ &= (s_{x,t-1} - s_{x,e-m-1}) - m^{-1} \sum_{t=e-m+1}^{e} (s_{x,t-1} - s_{x,e-m-1}) \end{aligned}$$

and

$$s_{x,t-1} - s_{x,e-m-1} = \sum_{j=e-m}^{t-1} \rho^{t-1-j} \epsilon_{x,j} + (\rho^{t-(e-m)} - 1) s_{x,e-m-1}. \tag{S.1}$$

Here, for $|\rho| < 1$, $s_{x,e-m-1}$ is strictly stationary and so therefore is $s_{x,t-1} - s_{x,e-m-1}$ since its summation term involves a only finite number of $\epsilon_{x,j}$. Strict stationarity of $x_{t-1} - \bar{x}_{-1}$ then follows, as does that of $\tau_{e,m}$. On setting $\rho = 1$ in (S.1),

$$s_{x,t-1} - s_{x,e-m-1} = \sum_{j=e-m}^{t-1} \epsilon_{x,j}$$

so that $x_{t-1} - \bar{x}_{-1}$ does not now depend on the unit root process $s_{x,e-m-1}$ and is therefore strictly stationary, along with $\tau_{e,m}$. In fact, we can extend our results to cases where ρ is allowed to be T-dependent such as occurs, for example, when the predictor is strongly persistent displaying either local or moderate deviations from a unit root. To this end, suppose that

$$\rho = 1 - cT^{-\theta}$$

for constants $c > 0$ and $\theta \in (0, 1]$. Here $\theta = 1$ corresponds to the local deviation case; $\theta < 1$ to the moderate deviation case. Then, expanding the term $\rho^{t-(e-m)}-1$ of (S.1) in powers of $T^{-\theta}$, we can write

$$\rho^{t-(e-m)}-1=-\{t-(e-m)\}cT^{-\theta}+o(T^{-\theta})$$

such that, for large T, we have approximately,

$$s_{x,t-1}-s_{x,e-m-1}=\sum_{j=e-m}^{t-1}\rho^{t-1-j}\epsilon_{x,j}-\{t-(e-m)\}cT^{-\theta}s_{x,e-m-1}.$$

Then, since $s_{x,e-m-1}=O_p(T^{\theta/2})$, we find

$$s_{x,t-1}-s_{x,e-m-1}=\sum_{j=e-m}^{t-1}\rho^{t-1-j}\epsilon_{x,j}+O_p(T^{-\theta/2})$$

such that, asymptotically, $s_{x,t-1}-s_{x,e-m-1}$ and $x_{t-1}-\bar{x}_{-1}$ do not depend on $s_{x,e-m-1}$. Strict stationarity of $\tau_{e,m}$ (for large T) then follows.

(ii) $\{\tau_{e,m}\}$ *is an* $m-1$ *dependent sequence.* This follows from the uncorrelatedness of $\epsilon_{y,t}$ and becomes evident on examining the numerator term of $\hat{b}$ in (6) which can be written as

$$\sum_{t=e-m+1}^{e}(x_{t-1}-\bar{x}_{-1})(y_t-\bar{y})=\sum_{t=e-m+1}^{e}(x_{t-1}-\bar{x}_{-1})\epsilon_{y,t}$$

from which we find that $\tau_{e,m}$ and $\tau_{e-k,m}$ are independent for $|k|>m-1$.

Property *(i)* (strict stationarity) and property *(ii)* (finite order dependence, which is equivalent to infinitely fast mixing) taken together are sufficient for us to be able to apply the result from Theorem 2.1 of Ferreira and Scotto (2002,p.478). Setting $r=s=1$ in their notation gives the limiting probability that the maximum of one disjoint subinterval exceeds the maximum of another as equal to the limiting ratio of the length of the former subinterval to the total length of the two subintervals. This result can then be used to establish (8) in Proposition 1.

Table S1. List of predictors used

Macroeconomic and financial variables (MFVs)

1. log dividend yield (dy_{t-1})
2. log dividend-price ratio (dp_{t-1})
3. log earnings-price ratio (ep_{t-1})
4. book-to-market ratio (bm_{t-1})
5. short term yield (st_{t-1})
6. long-term yield (lt_{t-1})
7. long-term - short-term yield spread ($sp_{t-1} = lt_{t-1} - st_{t-1}$)
8. BAA-AAA corporate bond yield spread (dsp_{t-1})
9. net equity expansion ($ntis_{t-1}$)
10. inflation (inf_{t-1})

Technical analysis indicators (TAIs)

1. 1-9 moving average rule ($MAI_{1,9,t-1}$)
2. 1-12 moving average rule indicator ($MAI_{1,12,t-1}$)
3. 2-9 moving average rule ($MAI_{2,9,t-1}$)
4. 2-12 moving average rule ($MAI_{2,12,t-1}$)
5. 9 period momentum rule ($MOI_{9,t-1}$)
6. 12 period momentum rule ($MOI_{12,t-1}$)
7. 1-9 on balance volume rule ($OBV_{1,9,t-1}$)
8. 1-12 on balance volume rule ($OBV_{1,12,t-1}$)
9. 2-9 on balance volume rule ($OBV_{2,9,t-1}$)
10. 2-12 on balance volume rule ($OBV_{2,12,t-1}$)

Table S2. Preliminary results for the full sample, 12/74-12/15

	$\hat{\beta}$	t_{NW}	IV_{comb}	$R^2(\%)$	$\bar{R}^2(\%)$
		MFVs			
dy_{t-1}	0.615	1.358*	0.840	0.395	0.191
dp_{t-1}	0.576	1.272	0.606	0.345	0.141
ep_{t-1}	0.424	0.721	1.072	0.225	0.022
bm_{t-1}	0.497	0.659	0.544	0.106	-0.098
st_{t-1}	0.042	0.723	0.350	0.116	-0.087
lt_{t-1}	0.036	0.501	0.398	0.056	-0.148
sp_{t-1}	0.108	0.804	-0.025	0.129	-0.075
dsp_{t-1}	0.135	0.214	0.064	0.021	-0.183
$ntis_{t-1}$	-0.005	-0.030	0.883	0.000	-0.204
inf_{t-1}	0.517	0.768	0.656	0.148	-0.056
		TAIs			
$MAI_{1,9,t-1}$	0.430	0.827	0.513	0.200	-0.004
$MAI_{1,12,t-1}$	0.647	1.103	1.142	0.415	0.211
$MAI_{2,9,t-1}$	0.453	0.855	1.126	0.215	0.012
$MAI_{2,12,t-1}$	0.802	1.465*	1.766*	0.648	0.445
$MOI_{9,t-1}$	0.370	0.621	0.209	0.136	-0.067
$MOI_{12,t-1}$	0.350	0.553	0.623	0.116	-0.088
$OBV_{1,9,t-1}$	0.491	0.994	0.045	0.269	0.065
$OBV_{1,12,t-1}$	0.679	1.265	0.150	0.488	0.285
$OBV_{2,9,t-1}$	0.759	1.483*	0.478	0.637	0.434
$OBV_{2,12,t-1}$	0.776	1.441*	0.761	0.642	0.439

Note. * denotes statistical significance at the 0.10 level. The critical value used for t_{NW} is 1.282. The critical value used for IV_{comb} is ± 1.645.

Table S3. MFVs: preliminary results for each training period used when monitoring with $m = \{15, 30, 60\}$

	$\hat{\beta}$	t_{NW}	IV_{comb}	$R^2(\%)$	$\bar{R}^2(\%)$
		$m = 15$			
dy_{t-1}	-0.010	-0.013	0.466	0.000	-0.352
dp_{t-1}	0.006	0.008	0.344	0.000	-0.352
ep_{t-1}	0.135	0.210	0.096	0.014	-0.338
bm_{t-1}	-0.142	-0.164	-0.183	0.010	-0.343
st_{t-1}	0.142	2.075*	2.384*	0.832	0.483
lt_{t-1}	0.165	1.467*	2.346*	0.634	0.284
sp_{t-1}	0.173	1.178	1.129	0.351	0.000
dsp_{t-1}	0.452	0.763	1.286	0.250	-0.101
$ntis_{t-1}$	0.362	2.748*	1.434	1.926	1.441
inf_{t-1}	1.307	1.913*	1.431	0.842	0.493
		$m = 30$			
dy_{t-1}	-0.100	-0.096	0.347	0.004	-0.367
dp_{t-1}	-0.125	-0.128	0.175	0.007	-0.365
ep_{t-1}	0.084	0.122	0.008	0.005	-0.367
bm_{t-1}	-0.189	-0.203	-0.135	0.016	-0.355
st_{t-1}	0.146	2.112*	2.466*	0.943	0.574
lt_{t-1}	0.181	1.548*	2.442*	0.747	0.378
sp_{t-1}	0.186	1.271	1.144	0.436	0.066
dsp_{t-1}	0.477	0.773	1.409	0.283	-0.087
$ntis_{t-1}$	0.362	2.748*	1.434	1.926	1.441
inf_{t-1}	1.283	1.828*	1.332	0.854	0.485
		$m = 60$			
dy_{t-1}	1.729	1.238	1.214	0.834	0.419
dp_{t-1}	1.721	1.328*	1.035	0.823	0.408
ep_{t-1}	0.569	0.794	0.632	0.226	-0.191
bm_{t-1}	0.854	0.800	0.675	0.272	-0.145
st_{t-1}	0.112	1.607*	2.405*	0.569	0.153
lt_{t-1}	0.108	0.837	2.200*	0.241	-0.177
sp_{t-1}	0.212	1.501*	1.027	0.602	0.186
dsp_{t-1}	1.085	1.714*	2.197*	1.316	0.903
$ntis_{t-1}$	0.362	2.748*	1.434	1.926	1.441
inf_{t-1}	0.928	1.298*	0.929	0.446	0.029

Note. * denotes statistical significance at the 0.10 level. The critical value used for t_{NW} is 1.282. The critical value used for IV_{comb} is ± 1.645. The training periods are 12/74-10/98 (for $m = 15$), 12/74-07/97 (for $m = 30$), and 12/74-01/95 (for $m = 60$) for all predictors other than $ntis_{t-1}$. For $ntis_{t-1}$ the training periods are 12/74-12/91 for all values of m.

Table S4. TAIs: preliminary results for each training period used when monitoring with $m = \{15, 30, 60\}$

	$\hat{\beta}$	t_{NW}	IV_{comb}	$R^2(\%)$	$\bar{R}^2(\%)$
		$m = 15$			
$MAI_{1,9,t-1}$	-0.718	-1.175	-0.969	0.537	0.187
$MAI_{1,12,t-1}$	-0.289	-0.473	-0.116	0.079	-0.273
$MAI_{2,9,t-1}$	-0.299	-0.524	-0.100	0.090	-0.262
$MAI_{2,12,t-1}$	0.061	0.113	0.520	0.003	-0.349
$MOI_{9,t-1}$	-0.249	-0.415	0.258	0.058	-0.294
$MOI_{12,t-1}$	-0.464	-0.715	0.256	0.184	-0.167
$OBV_{1,9,t-1}$	0.264	0.472	0.659	0.072	-0.280
$OBV_{1,12,t-1}$	0.200	0.301	0.362	0.037	-0.315
$OBV_{2,9,t-1}$	0.380	0.600	0.648	0.141	-0.210
$OBV_{2,12,t-1}$	0.209	0.309	0.700	0.041	-0.311
		$m = 30$			
$MAI_{1,9,t-1}$	-0.528	-0.893	-0.693	0.309	-0.062
$MAI_{1,12,t-1}$	-0.060	-0.104	0.236	0.004	-0.368
$MAI_{2,9,t-1}$	-0.085	-0.158	0.218	0.008	-0.364
$MAI_{2,12,t-1}$	0.201	0.379	0.819	0.040	-0.331
$MOI_{9,t-1}$	-0.250	-0.415	0.249	0.064	-0.308
$MOI_{12,t-1}$	-0.468	-0.715	0.248	0.204	-0.167
$OBV_{1,9,t-1}$	0.367	0.660	0.729	0.150	-0.222
$OBV_{1,12,t-1}$	0.206	0.308	0.360	0.043	-0.329
$OBV_{2,9,t-1}$	0.522	0.830	0.934	0.286	-0.084
$OBV_{2,12,t-1}$	0.215	0.316	0.707	0.048	-0.324
		$m = 60$			
$MAI_{1,9,t-1}$	-0.792	-1.316	-1.089	0.703	0.288
$MAI_{1,12,t-1}$	-0.302	-0.513	-0.101	0.094	-0.324
$MAI_{2,9,t-1}$	-0.251	-0.458	0.016	0.068	-0.350
$MAI_{2,12,t-1}$	-0.030	-0.056	0.562	0.001	-0.417
$MOI_{9,t-1}$	-0.502	-0.829	-0.039	0.265	-0.153
$MOI_{12,t-1}$	-0.659	-0.980	-0.049	0.414	-0.003
$OBV_{1,9,t-1}$	0.130	0.233	0.460	0.019	-0.399
$OBV_{1,12,t-1}$	-0.026	-0.039	0.033	0.001	-0.418
$OBV_{2,9,t-1}$	0.299	0.467	0.643	0.096	-0.322
$OBV_{2,12,t-1}$	-0.019	-0.028	0.402	0.000	-0.418

Note. * denotes statistical significance at the 0.01 level. The critical value used for t_{NW} is 1.282. The critical value used for IV_{comb} is ± 1.645. The training periods are 12/74-10/98 (for $m = 15$), 12/74-07/97 (for $m = 30$), and 12/74-01/95 (for $m = 60$).

Table S5. Number of predictive regimes detected by SEQ with $\pi = 0.05$

	$m = 15$	$m = 30$	$m = 60$
	MFVs		
dy_{t-1}	1	0	2
dp_{t-1}	1	1	2
ep_{t-1}	1	0	0
bm_{t-1}	1	0	1
st_{t-1}	0	0	0
lt_{t-1}	3	1	1
sp_{t-1}	0	0	0
dsp_{t-1}	1	0	0
$ntis_{t-1}$	0	0	0
inf_{t-1}	0	0	0
	TAIs		
$MAI_{1,9,t-1}$	0	1	1
$MAI_{1,12,t-1}$	0	2	2
$MAI_{2,9,t-1}$	0	0	2
$MAI_{2,12,t-1}$	0	1	3
$MOI_{9,t-1}$	1	3	3
$MOI_{12,t-1}$	0	3	2
$OBV_{1,9,t-1}$	0	1	1
$OBV_{1,12,t-1}$	0	2	3
$OBV_{2,9,t-1}$	0	0	0
$OBV_{2,12,t-1}$	0	1	3

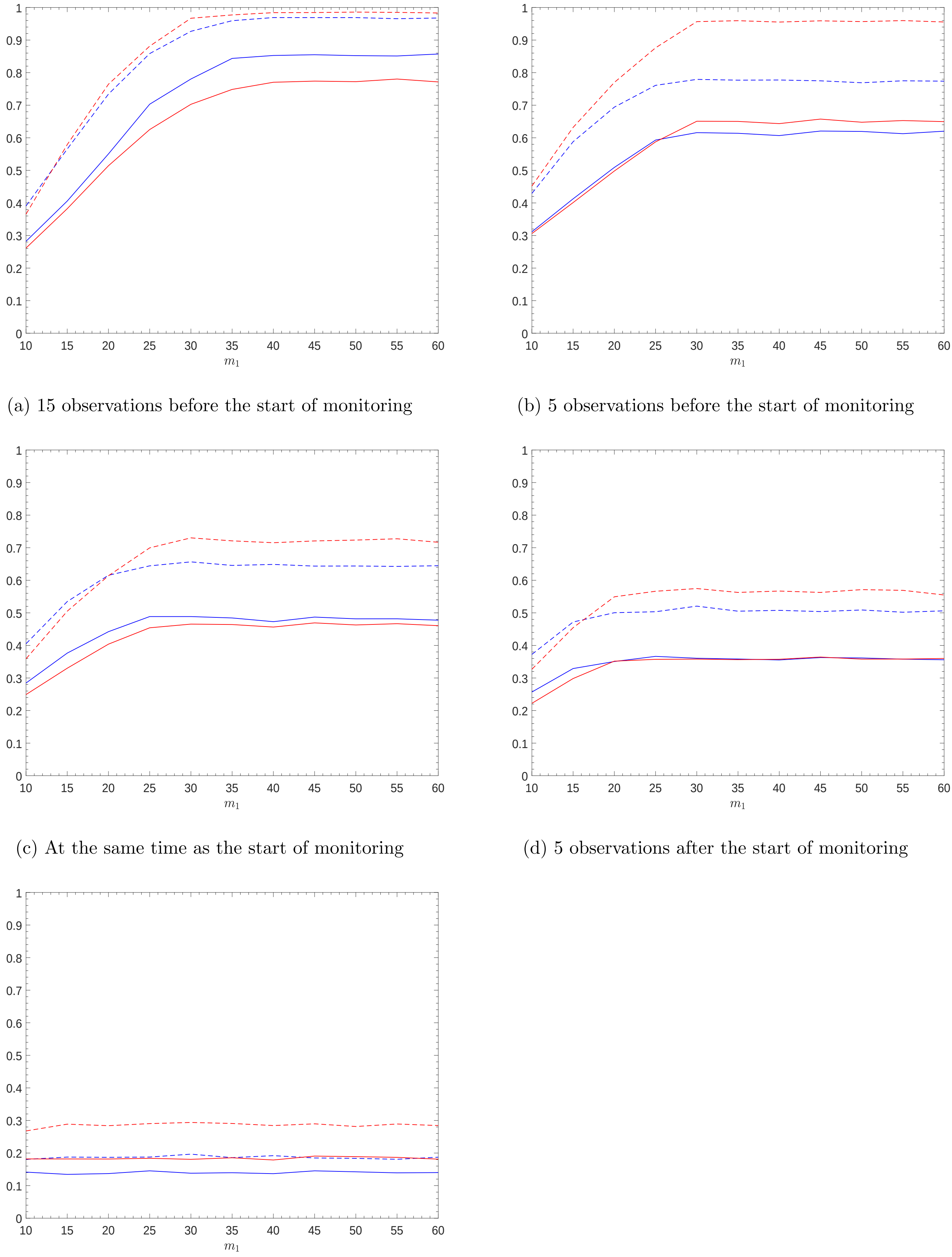
Table S6. First month where a predictive regime is detected by SEQ with $\pi = 0.05$

	$m = 15$		$m = 30$		$m = 60$	
	SEQ	FPR_{SEQ}	SEQ	FPR_{SEQ}	SEQ	FPR_{SEQ}
			MFVs			
dy_{t-1}	07/07	0.251	N/A	N/A	04/02	0.133
dp_{t-1}	05/15	0.405	05/01	0.066	10/01	0.108
ep_{t-1}	10/14	0.396	N/A	N/A	N/A	N/A
bm_{t-1}	10/00	0.035	N/A	N/A	10/01	0.108
st_{t-1}	N/A	N/A	N/A	N/A	N/A	N/A
lt_{t-1}	08/03	0.139	08/03	0.154	06/05	0.266
sp_{t-1}	N/A	N/A	N/A	N/A	N/A	N/A
dsp_{t-1}	06/12	0.355	N/A	N/A	N/A	N/A
$ntis_{t-1}$	N/A	N/A	N/A	N/A	N/A	N/A
inf_{t-1}	N/A	N/A	N/A	N/A	N/A	N/A
			TAIs			
$MAI_{1,9,t-1}$	N/A	N/A	04/10	0.339	03/09	0.379
$MAI_{1,12,t-1}$	N/A	N/A	09/03	0.157	04/04	0.222
$MAI_{2,9,t-1}$	N/A	N/A	N/A	N/A	12/08	0.372
$MAI_{2,12,t-1}$	N/A	N/A	04/09	0.316	10/02	0.157
$MOI_{9,t-1}$	07/03	0.137	09/03	0.157	08/01	0.099
$MOI_{12,t-1}$	N/A	N/A	02/04	0.171	08/01	0.099
$OBV_{1,9,t-1}$	N/A	N/A	04/10	0.339	02/10	0.401
$OBV_{1,12,t-1}$	N/A	N/A	01/09	0.311	04/10	0.405
$OBV_{2,9,t-1}$	N/A	N/A	N/A	N/A	N/A	N/A
$OBV_{2,12,t-1}$	N/A	N/A	01/09	0.311	04/05	0.260

Table S7. Out-of-sample forecasting results for MFVs with $m = 30$

	$MSFE_B$	$MSFE_{PR}$	p-value	OS-R^2(%)
		MAX		
dy_{t-1}	32.790	31.058	0.260	5.282
dp_{t-1}	32.790	31.223	0.807	1.730
ep_{t-1}	16.495	16.421	0.684	0.447
bm_{t-1}	N/A	N/A	N/A	N/A
st_{t-1}	22.785	22.748	0.321	0.166
lt_{t-1}	23.310	23.650	0.505	-1.460
sp_{t-1}	N/A	N/A	N/A	N/A
dsp_{t-1}	22.760	22.782	0.321	-0.096
$ntis_{t-1}$	23.294	23.338	0.179	-0.187
inf_{t-1}	N/A	N/A	N/A	N/A
		SEQ		
dy_{t-1}	N/A	N/A	N/A	N/A
dp_{t-1}	N/A	N/A	N/A	N/A
ep_{t-1}	26.248	27.003	0.129	-2.878
bm_{t-1}	N/A	N/A	N/A	N/A
st_{t-1}	N/A	N/A	N/A	N/A
lt_{t-1}	23.088	23.091	0.985	-0.016
sp_{t-1}	N/A	N/A	N/A	N/A
dsp_{t-1}	N/A	N/A	N/A	N/A
$ntis_{t-1}$	N/A	N/A	N/A	N/A
inf_{t-1}	N/A	N/A	N/A	N/A

Note. $MSFE_B$ is the mean squared forecast error for a fixed mean benchmark model; $MSFE_{PR}$ is the mean squared forecast error allowing for a single predictability regime detected using MAX; p-value is the statistical significance of the Diebold and Mariano (1995) test for equal forcasting accuracy (employing the Harvey *et al.*, 1997; bias-correction and Student's t-critical values), OS-R^2(%) is the out-of-sample R^2 value.



(a) 15 observations before the start of monitoring

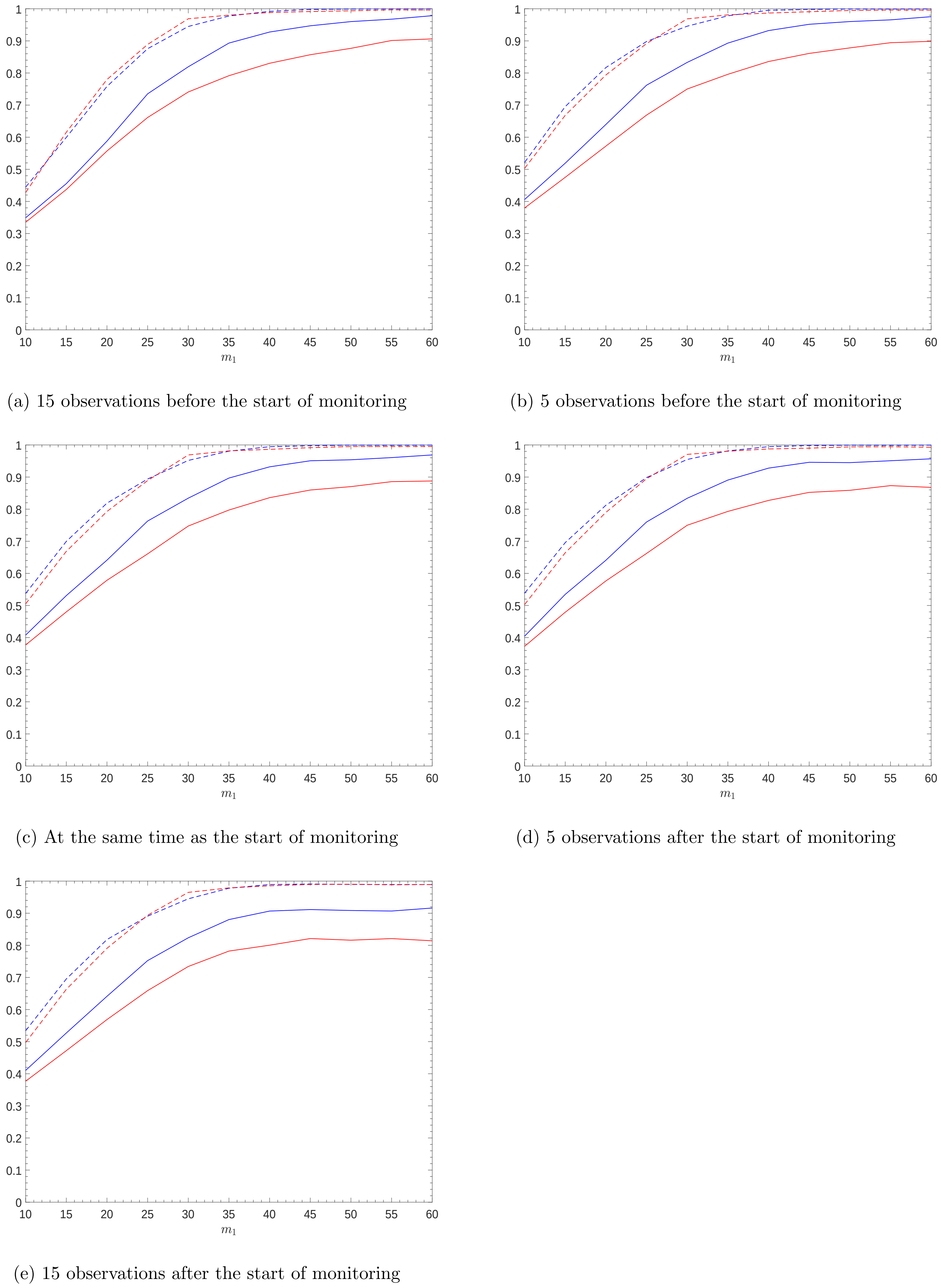
(b) 5 observations before the start of monitoring

(c) At the same time as the start of monitoring

(d) 5 observations after the start of monitoring

(e) 15 observations after the start of monitoring

Figure S1. Predictive regime detection frequency as a function of m_1 for different values of β_1: $T^* + m = 302$, $E = 327$, $\rho = 0.965$, $m = 30$; MAX, $\beta_1 = 0.25$, —; MAX, $\beta_1 = 0.50$, - - -; SEQ, $\beta_1 = 0.25$, —; SEQ, $\beta_1 = 0.50$, - - -



(a) 15 observations before the start of monitoring

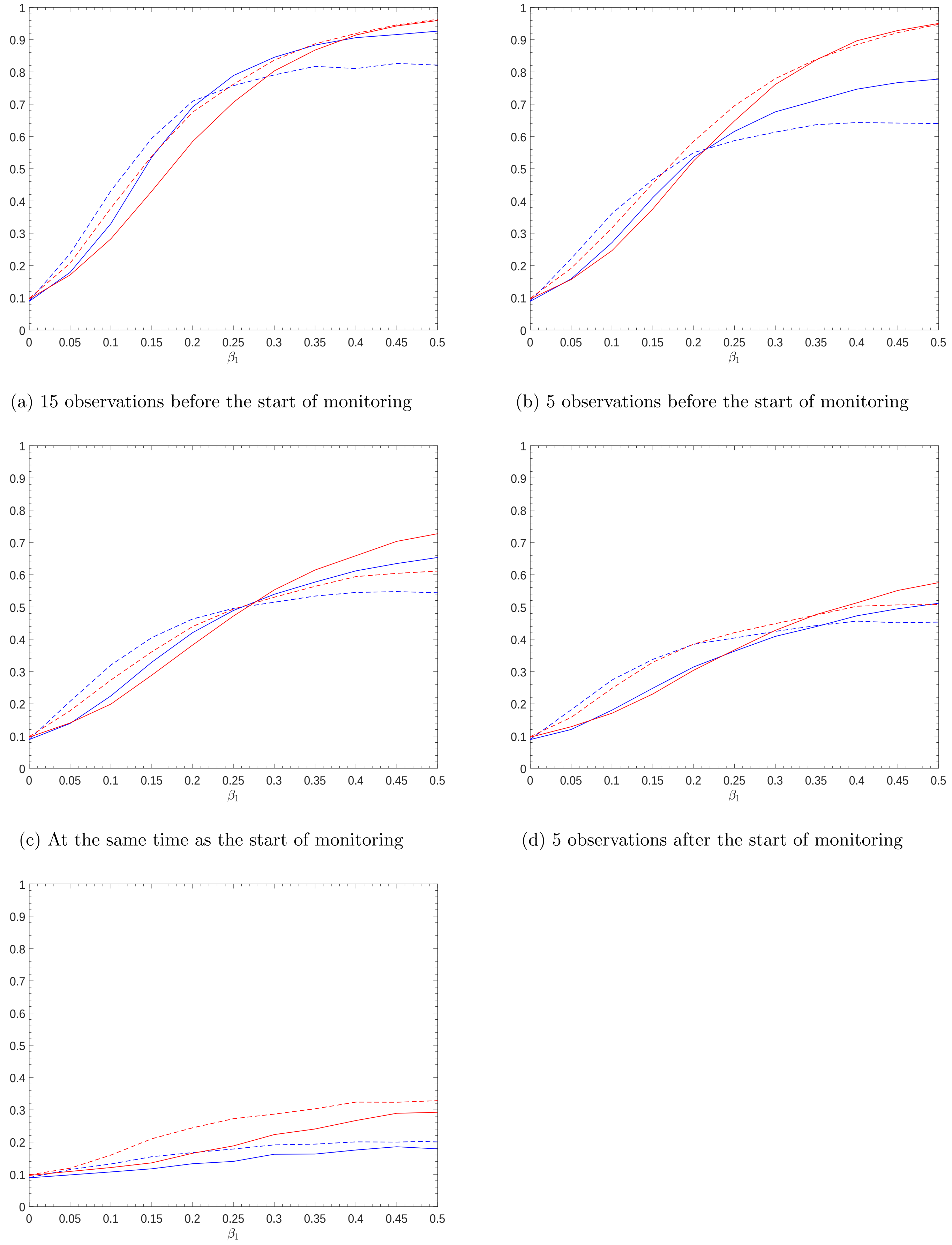
(b) 5 observations before the start of monitoring

(c) At the same time as the start of monitoring

(d) 5 observations after the start of monitoring

(e) 15 observations after the start of monitoring

Figure S2. Predictive regime detection frequency as a function of m_1 for different values of β_1: $T^* + m = 302$, $E = 361$, $\rho = 0.965$, $m = 30$; MAX, $\beta_1 = 0.25$, —; MAX, $\beta_1 = 0.50$, - - -; SEQ, $\beta_1 = 0.25$, —; SEQ, $\beta_1 = 0.50$, - - -



(a) 15 observations before the start of monitoring

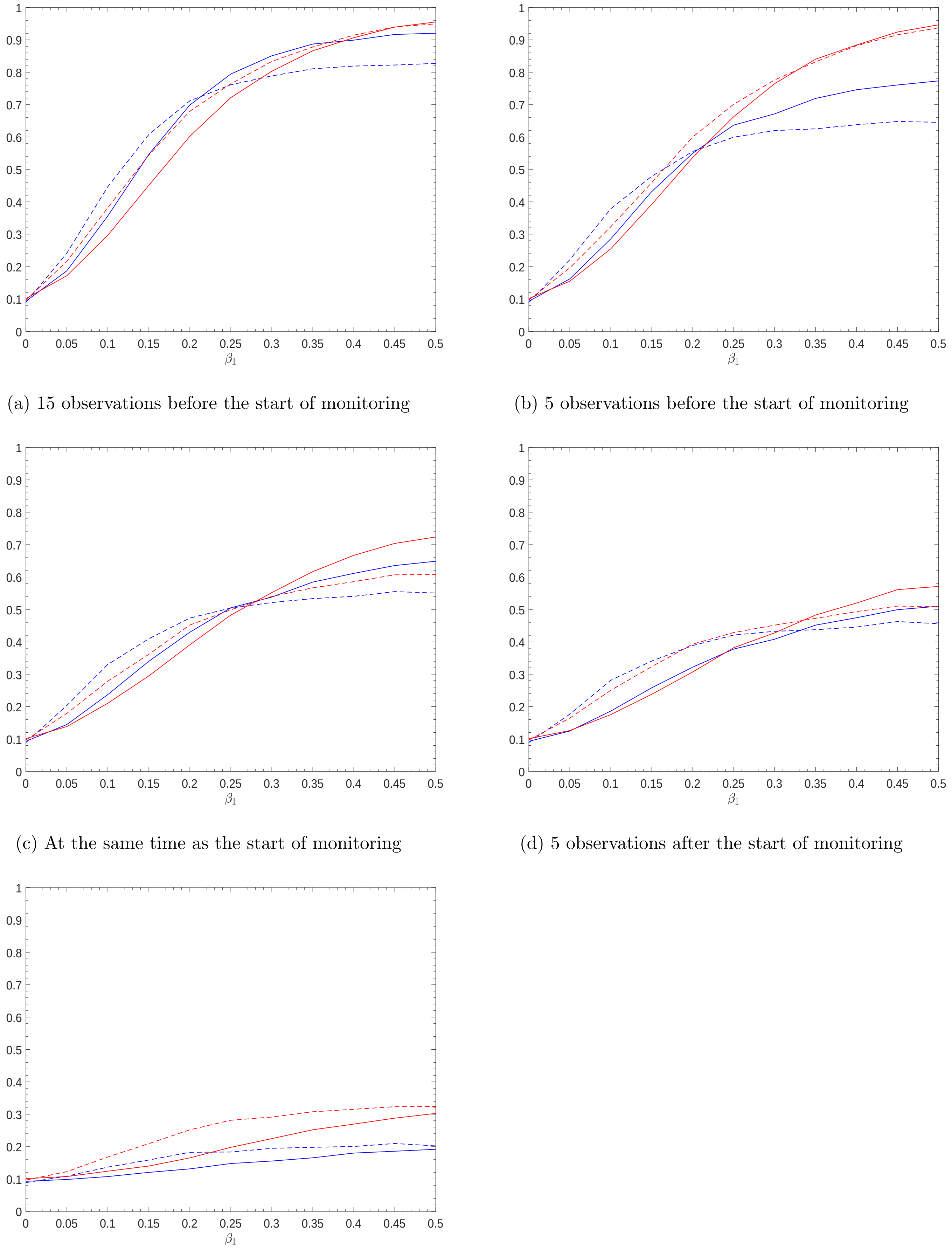
(b) 5 observations before the start of monitoring

(c) At the same time as the start of monitoring

(d) 5 observations after the start of monitoring

(e) 15 observations after the start of monitoring

Figure S3. Predictive regime detection frequency as a function of β_1 for different values of ρ and $t(10)$ error terms: $T^* + m = 302$, $E = 327$, $m_1 = 30$, $m = 30$; MAX, $\rho = 0.965$, —; MAX, $\rho = 0.995$, - - -; SEQ, $\rho = 0.965$, —; SEQ, $\rho = 0.995$, - - -



(a) 15 observations before the start of monitoring

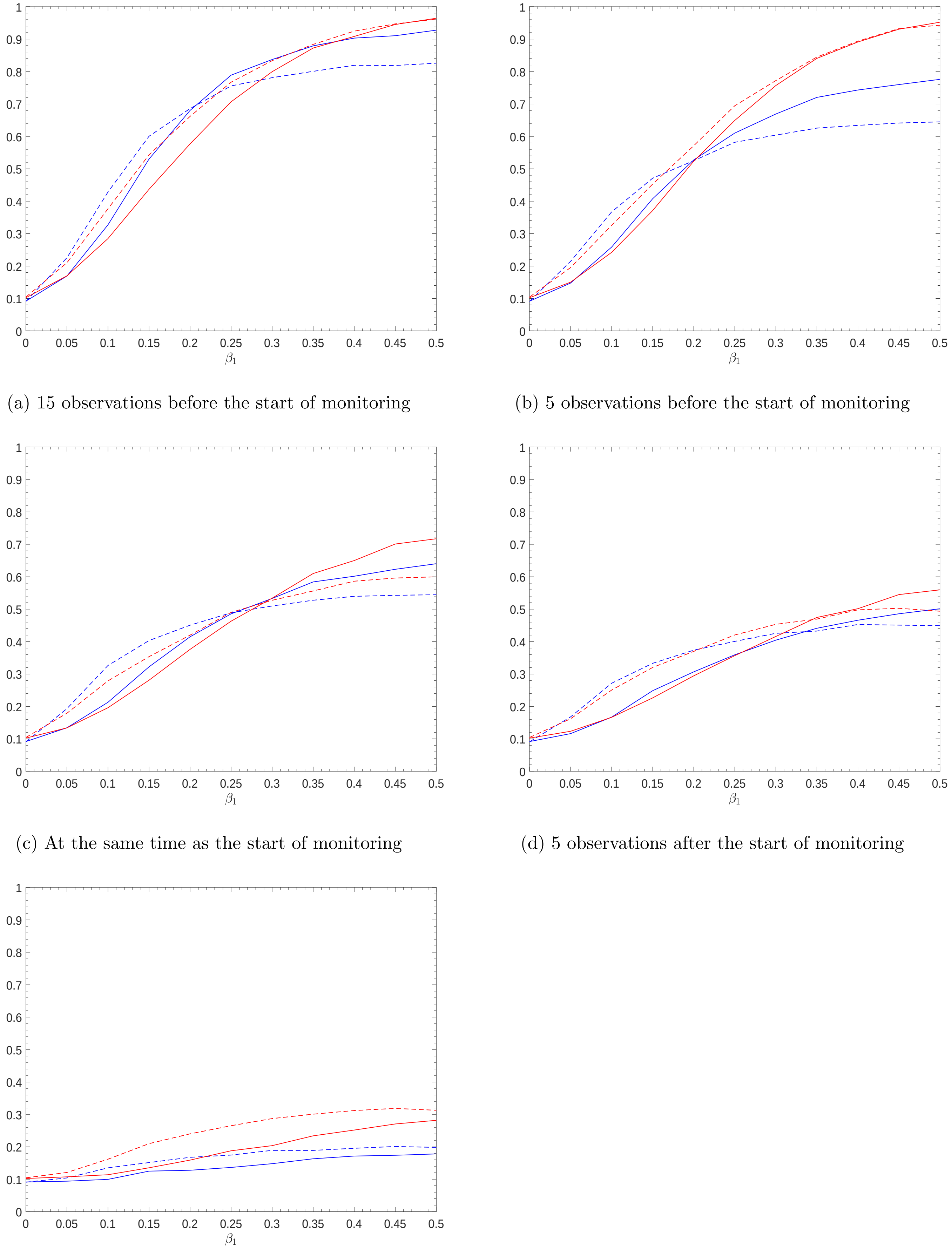
(b) 5 observations before the start of monitoring

(c) At the same time as the start of monitoring

(d) 5 observations after the start of monitoring

(e) 15 observations after the start of monitoring

Figure S4. Predictive regime detection frequency as a function of β_1 for different values of ρ and $t(5)$ error terms: $T^* + m = 302$, $E = 327$, $m_1 = 30$, $m = 30$; MAX, $\rho = 0.965$, —; MAX, $\rho = 0.995$, - - -; SEQ, $\rho = 0.965$, —; SEQ, $\rho = 0.995$, - - -



(a) 15 observations before the start of monitoring

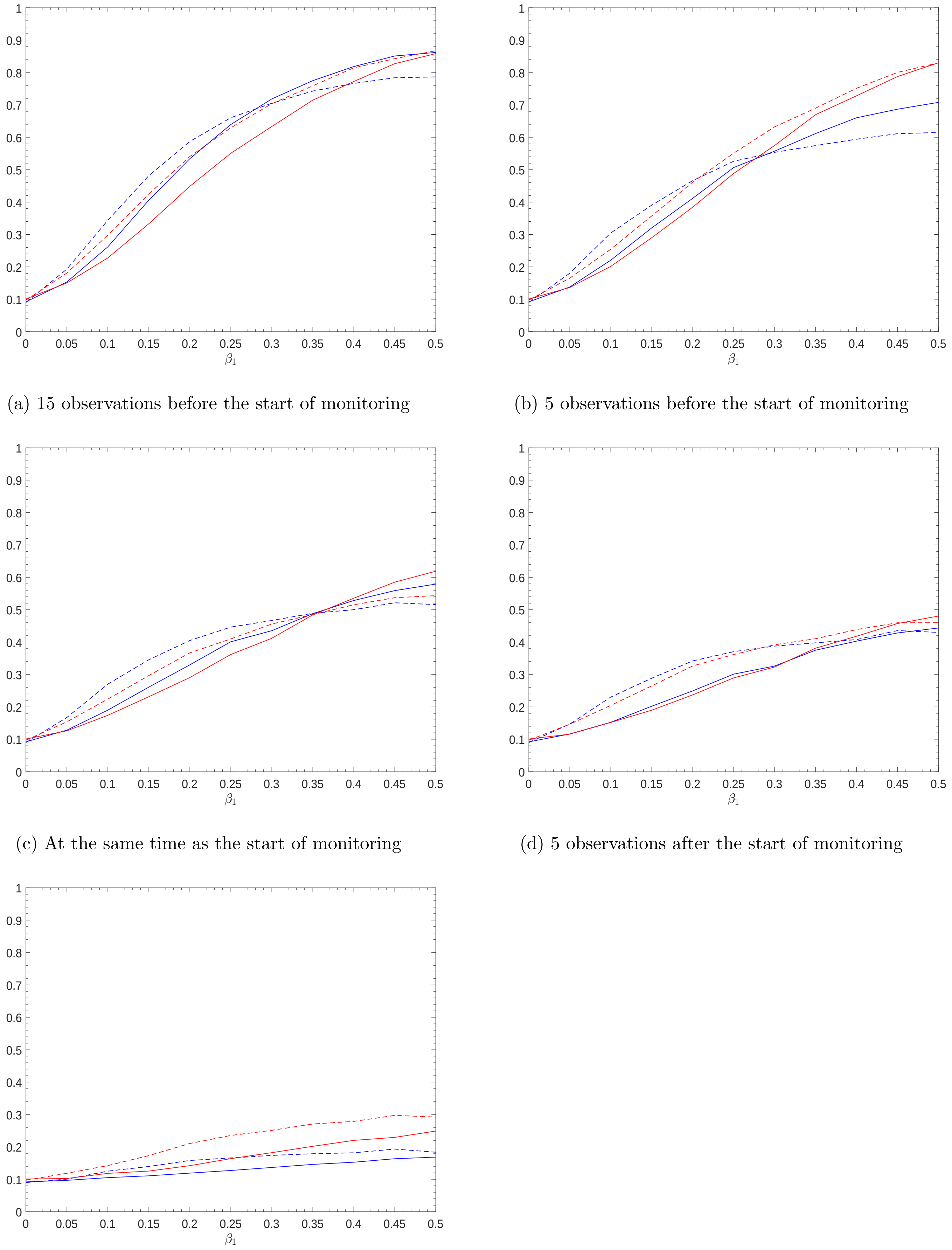
(b) 5 observations before the start of monitoring

(c) At the same time as the start of monitoring

(d) 5 observations after the start of monitoring

(e) 15 observations after the start of monitoring

Figure S5. Predictive regime detection frequency as a function of β_1 for different values of ρ and normally distributed GARCH error terms: $T^* + m = 302$, $E = 327$, $m_1 = 30$, $m = 30$; MAX, $\rho = 0.965$, —; MAX, $\rho = 0.995$, - - -; SEQ, $\rho = 0.965$, —; SEQ, $\rho = 0.995$, - - -



(a) 15 observations before the start of monitoring

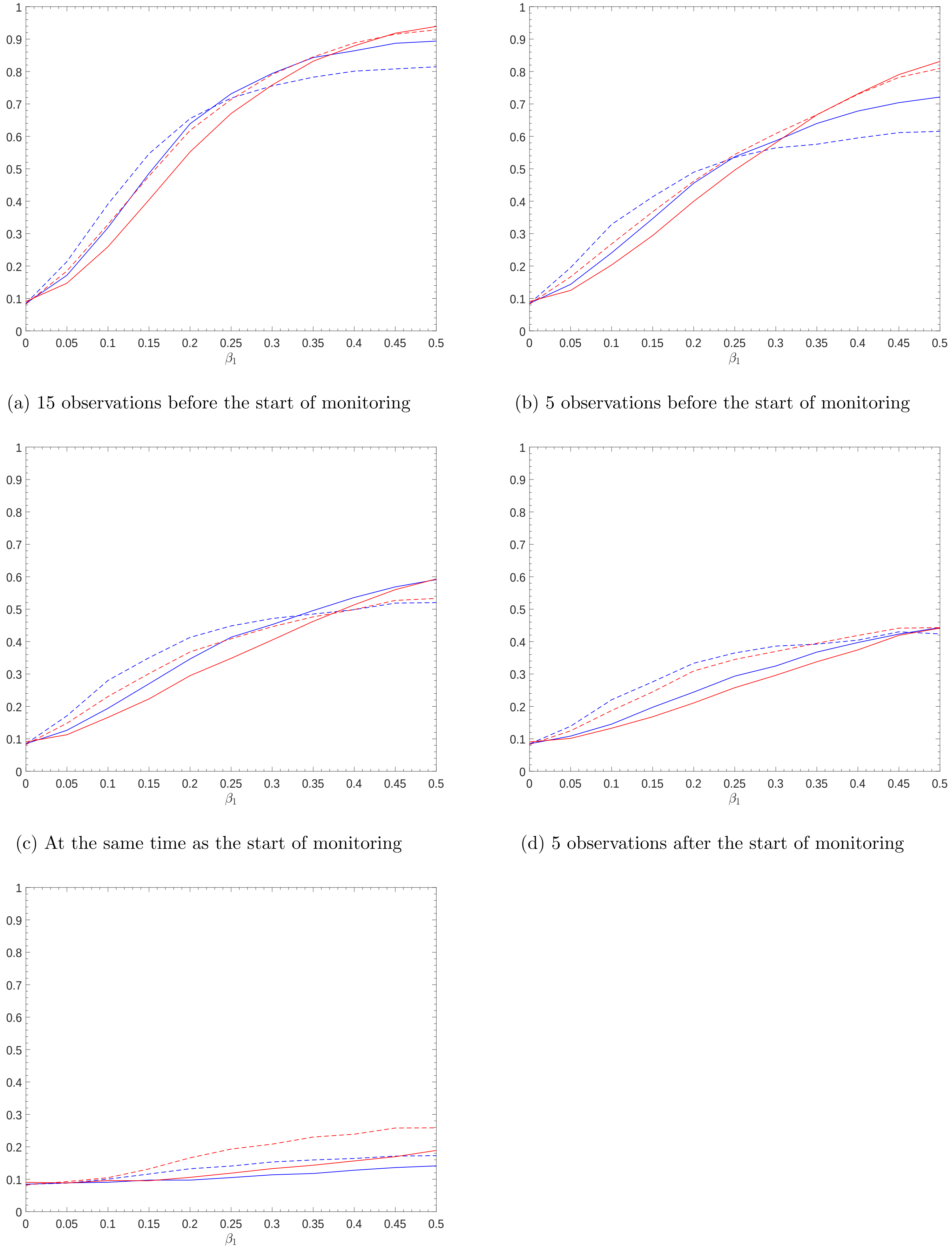
(b) 5 observations before the start of monitoring

(c) At the same time as the start of monitoring

(d) 5 observations after the start of monitoring

(e) 15 observations after the start of monitoring

Figure S6. Predictive regime detection frequency as a function of β_1 for different values of ρ and $t(5)$-GARCH error terms: $T^* + m = 302$, $E = 327$, $m_1 = 30$, $m = 30$; MAX, $\rho = 0.965$, —; MAX, $\rho = 0.995$, - - -; SEQ, $\rho = 0.965$, —; SEQ, $\rho = 0.995$, - - -



(a) 15 observations before the start of monitoring

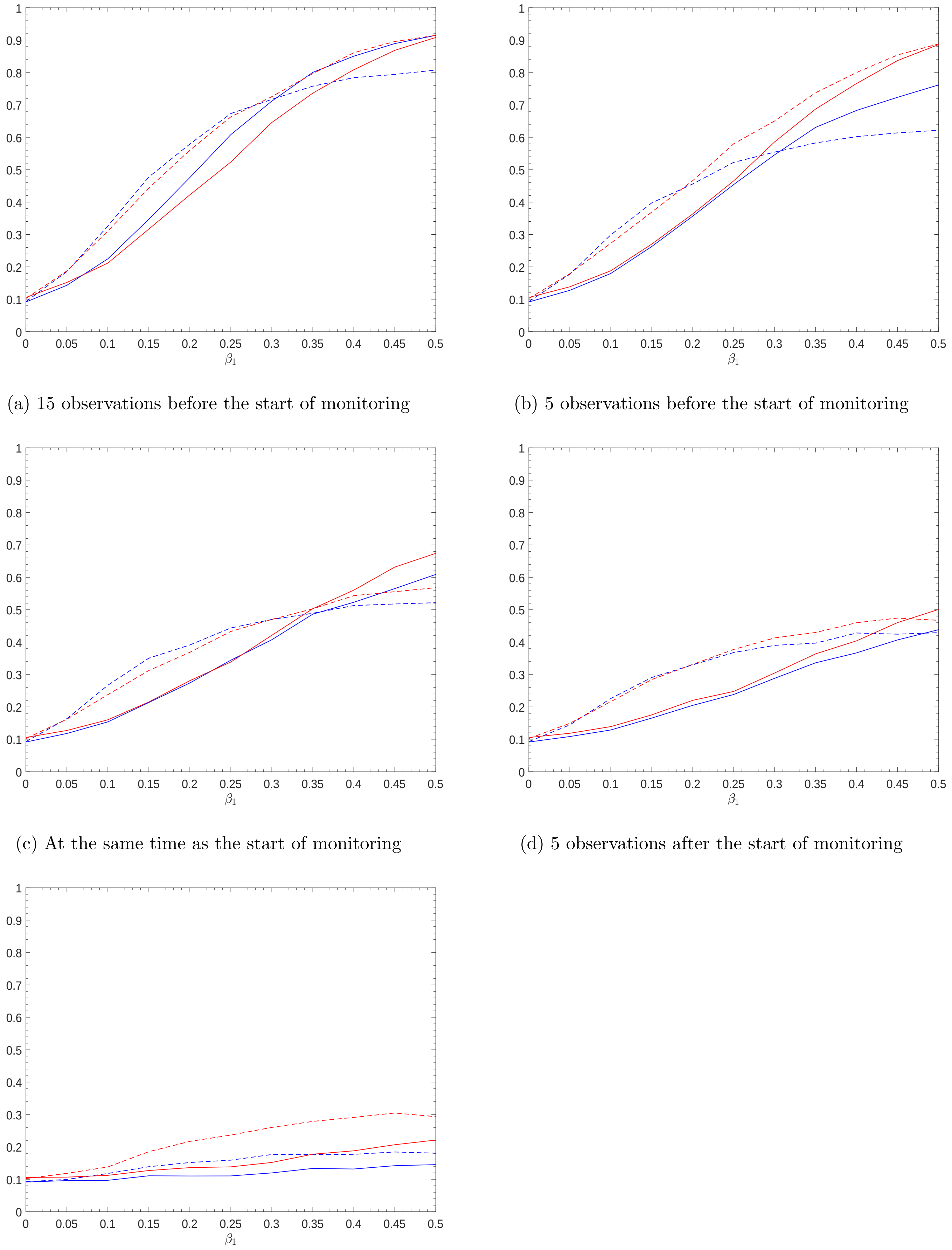
(b) 5 observations before the start of monitoring

(c) At the same time as the start of monitoring

(d) 5 observations after the start of monitoring

(e) 15 observations after the start of monitoring

Figure S7. Predictive regime detection frequency as a function of β_1 for different values of ρ and $t(5)$ errors with an unconditional volatility shift at $t = 315$ from $\sigma_y = 1$ to $\sigma_y = 2$: $T^* + m = 302$, $E = 327$, $m_1 = 30$, $m = 30$; MAX, $\rho = 0.965$, —; MAX, $\rho = 0.995$, - - -; SEQ, $\rho = 0.965$, —; SEQ, $\rho = 0.995$, - - -



(a) 15 observations before the start of monitoring

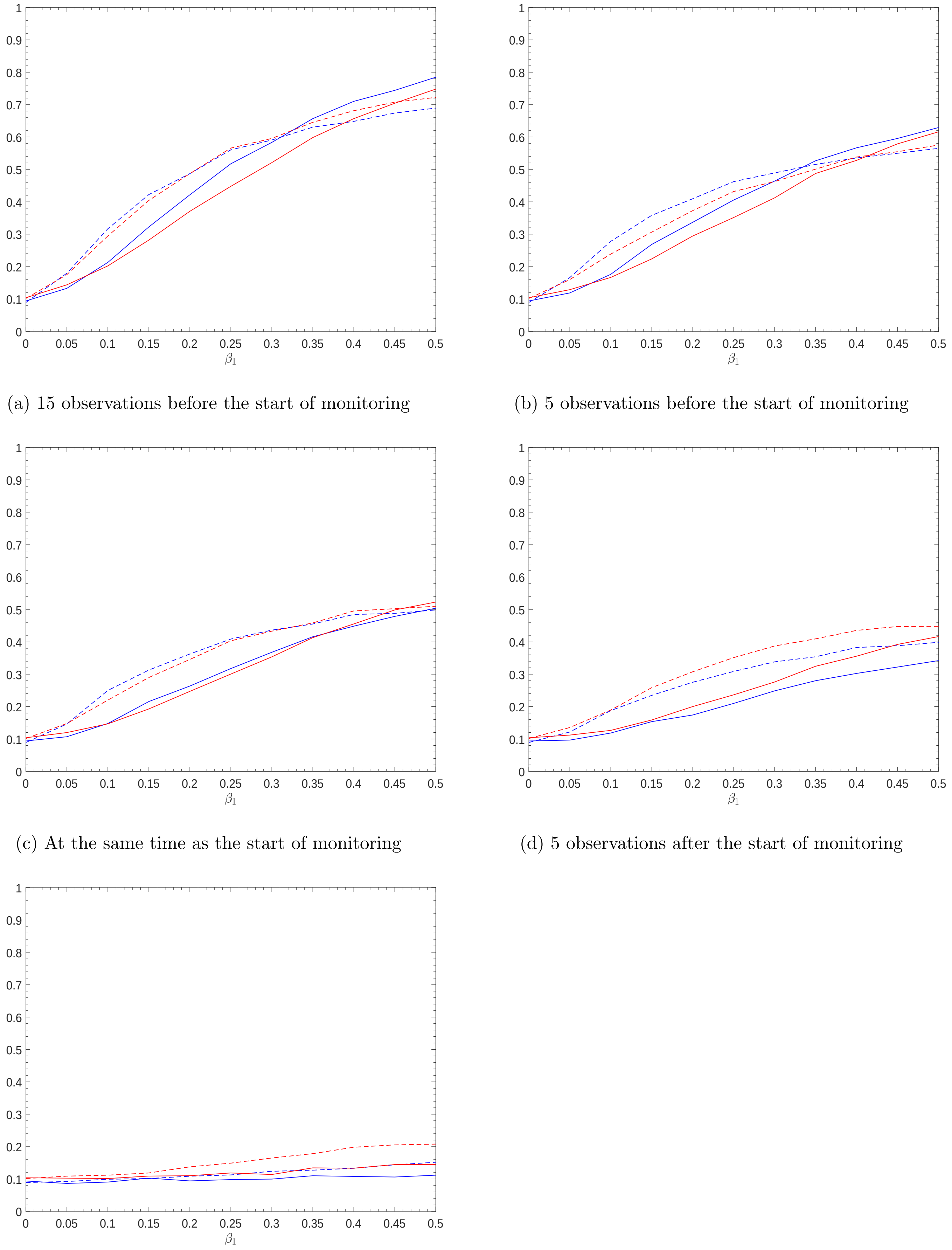
(b) 5 observations before the start of monitoring

(c) At the same time as the start of monitoring

(d) 5 observations after the start of monitoring

(e) 15 observations after the start of monitoring

Figure S8. Predictive regime detection frequency as a function of β_1 with an AR(2) predictor and AR parameters $\rho_1 = 0.595$, $\rho_2 = \{0.30, 0.40\}$: $T^* + m = 302$, $E = 327$, $m_1 = 30$, $m = 30$; MAX, $\rho_2 = 0.30$, —; MAX, $\rho_2 = 0.40$, - - -; SEQ, $\rho_2 = 0.30$, —; SEQ, $\rho_2 = 0.40$, - - -



(a) 15 observations before the start of monitoring

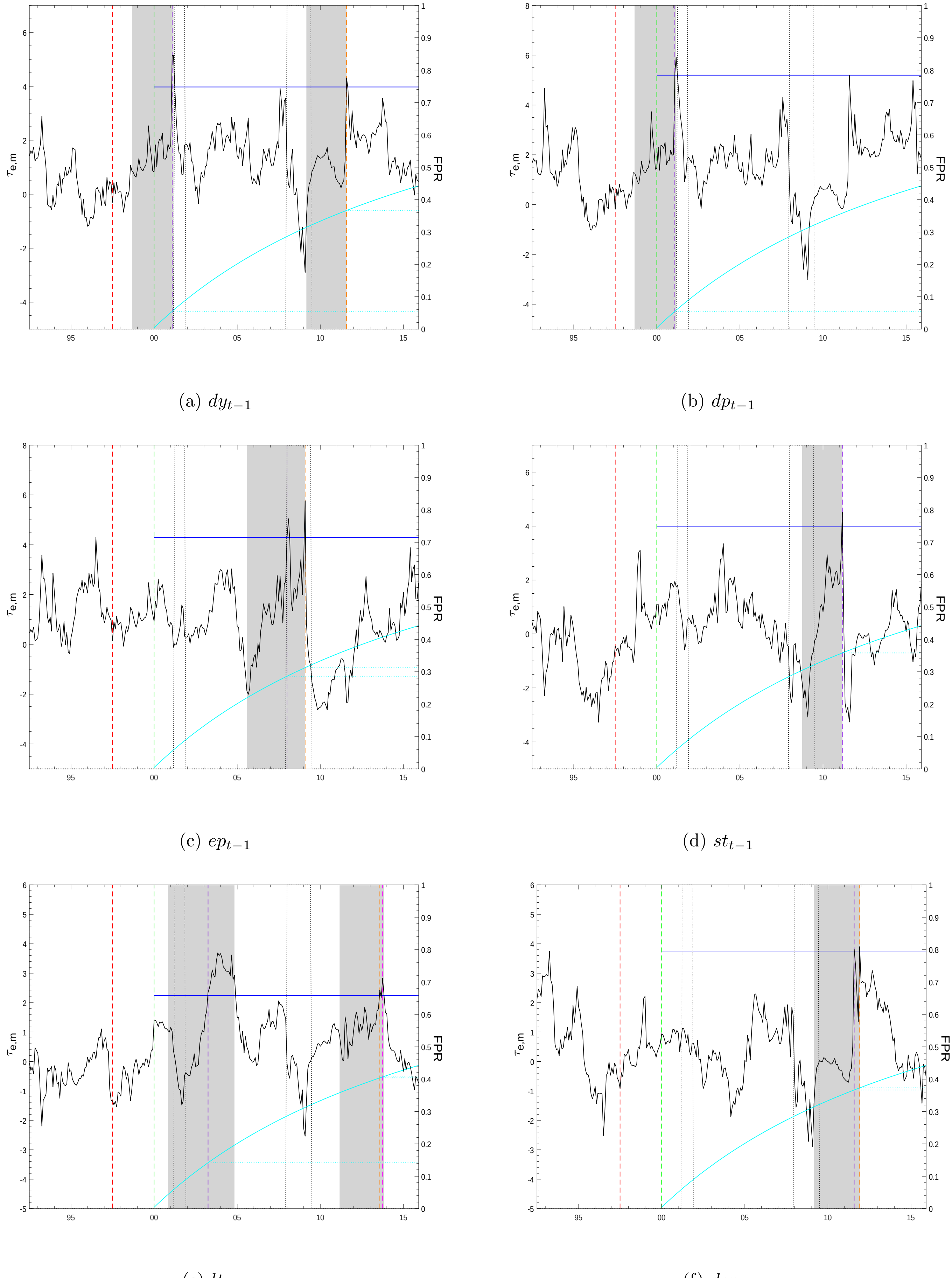
(b) 5 observations before the start of monitoring

(c) At the same time as the start of monitoring

(d) 5 observations after the start of monitoring

(e) 15 observations after the start of monitoring

Figure S9. Predictive regime detection frequency as a function of β_1 for different values of ρ and smooth regime change: $T^* + m = 302$, $E = 327$, $m = 30$; MAX, $\rho = 0.965$, —; MAX, $\rho = 0.995$, - - -; SEQ, $\rho = 0.965$, —; SEQ, $\rho = 0.995$, - - -



(a) dy_{t-1}

(b) dp_{t-1}

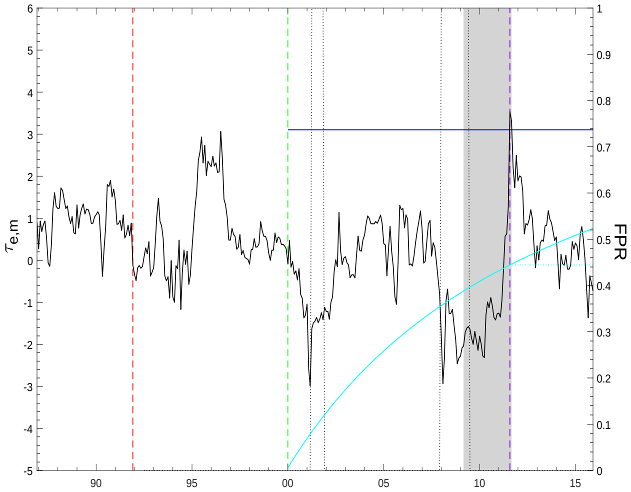
(c) ep_{t-1}

(d) st_{t-1}

(e) lt_{t-1}

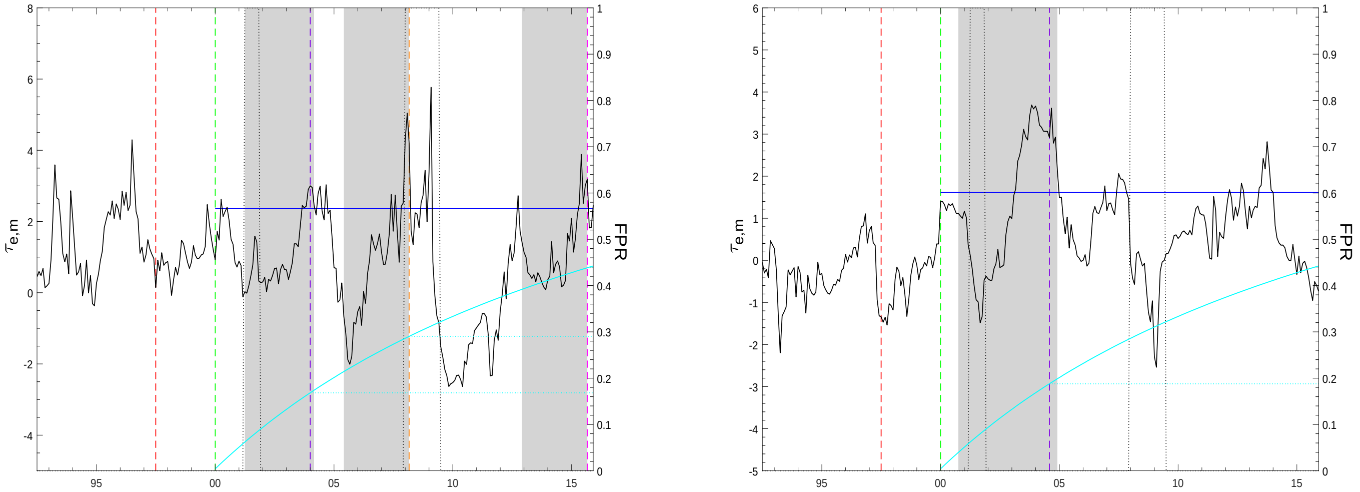
(f) dsp_{t-1}

Figure S10. MAX procedure, $m = 30$: —($\tau_{e,m}$), —($\max_{e\in[m+1,T^*]} \tau_{e,m}$), - - -($T^*$), - - -($T^* + m$), - - -(first rejection), - - -(second rejection), - - -(third rejection), ■ (weak set of dates), —(false positive rate), ⋯⋯(NBER indicator)



(g) $ntis_{t-1}$

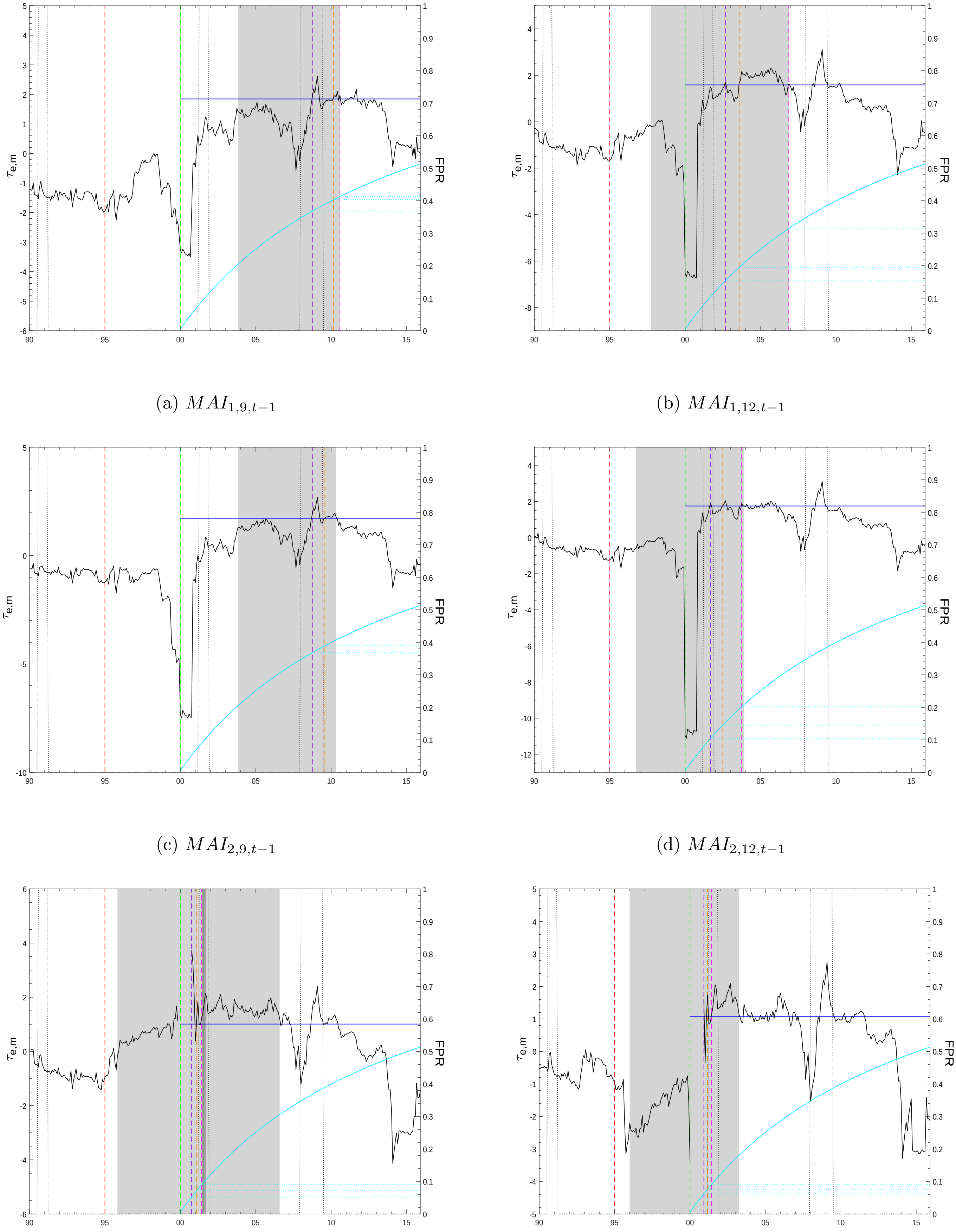
Figure S10 Continued. MAX procedure, $m = 30$: —($\tau_{e,m}$), —($\max_{e \in [m+1,T^*]} \tau_{e,m}$), - - -($T^*$), - - -($T^* + m$), - - -(first rejection), - - -(second rejection), - - -(third rejection), ▨ (weak set of dates), —(false positive rate), ·····(NBER indicator)



(a) ep_{t-1}

(b) lt_{t-1}

Figure S11. SEQ procedure, $m = 30$: —($\tau_{e,m}$), —($cv_{0.10}$), - - -(T^*), - - -(T^*+m), - - -(first rejection), - - -(second rejection), - - -(third rejection), ▨ (weak set of dates), —(false positive rate), ·····(NBER indicator)



(a) $MAI_{1,9,t-1}$

(b) $MAI_{1,12,t-1}$

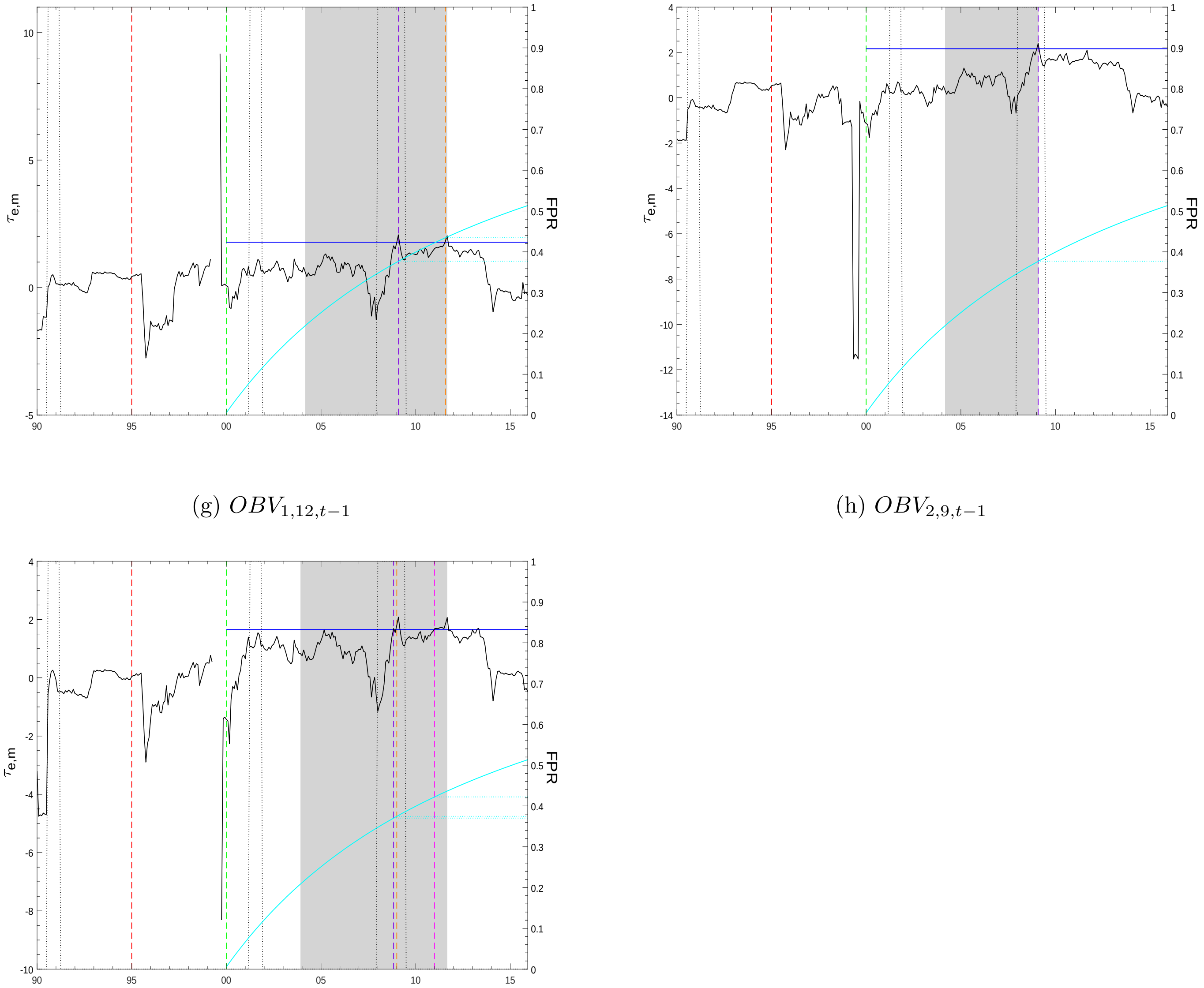
(c) $MAI_{2,9,t-1}$

(d) $MAI_{2,12,t-1}$

(e) $MOI_{9,t-1}$

(f) $MOI_{12,t-1}$

Figure S12. MAX procedure, $m = 60$: —($\tau_{e,m}$), —($\max_{e\in[m+1,T^*]}\tau_{e,m}$), - - -($T^*$), - - -($T^* + m$), - - -(first rejection), - - -(second rejection), - - -(third rejection), ▨ (weak set of dates), ■ (strong set of dates), —(false positive rate), ⋯⋯(NBER indicator)

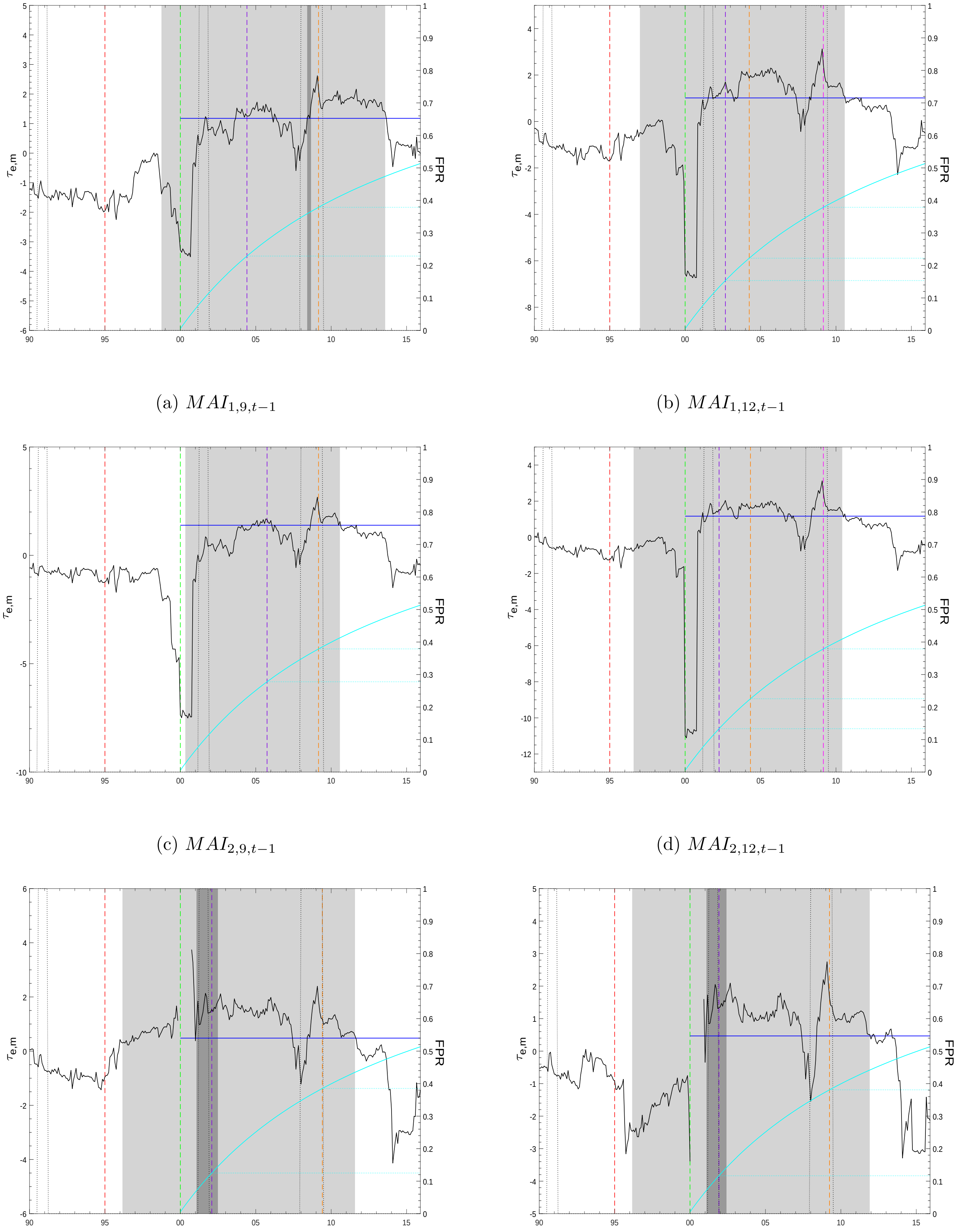


(g) $OBV_{1,12,t-1}$

(h) $OBV_{2,9,t-1}$

(i) $OBV_{2,12,t-1}$

Figure S12 Continued. MAX procedure, $m = 60$: —($\tau_{e,m}$), —($\max_{e\in[m+1,T^*]} \tau_{e,m}$), - - -($T^*$), - - -($T^* + m$), - - -(first rejection), - - -(second rejection), - - -(third rejection), ▒ (weak set of dates), ▓ (strong set of dates), —(false positive rate), ·····(NBER indicator)



(a) $MAI_{1,9,t-1}$

(b) $MAI_{1,12,t-1}$

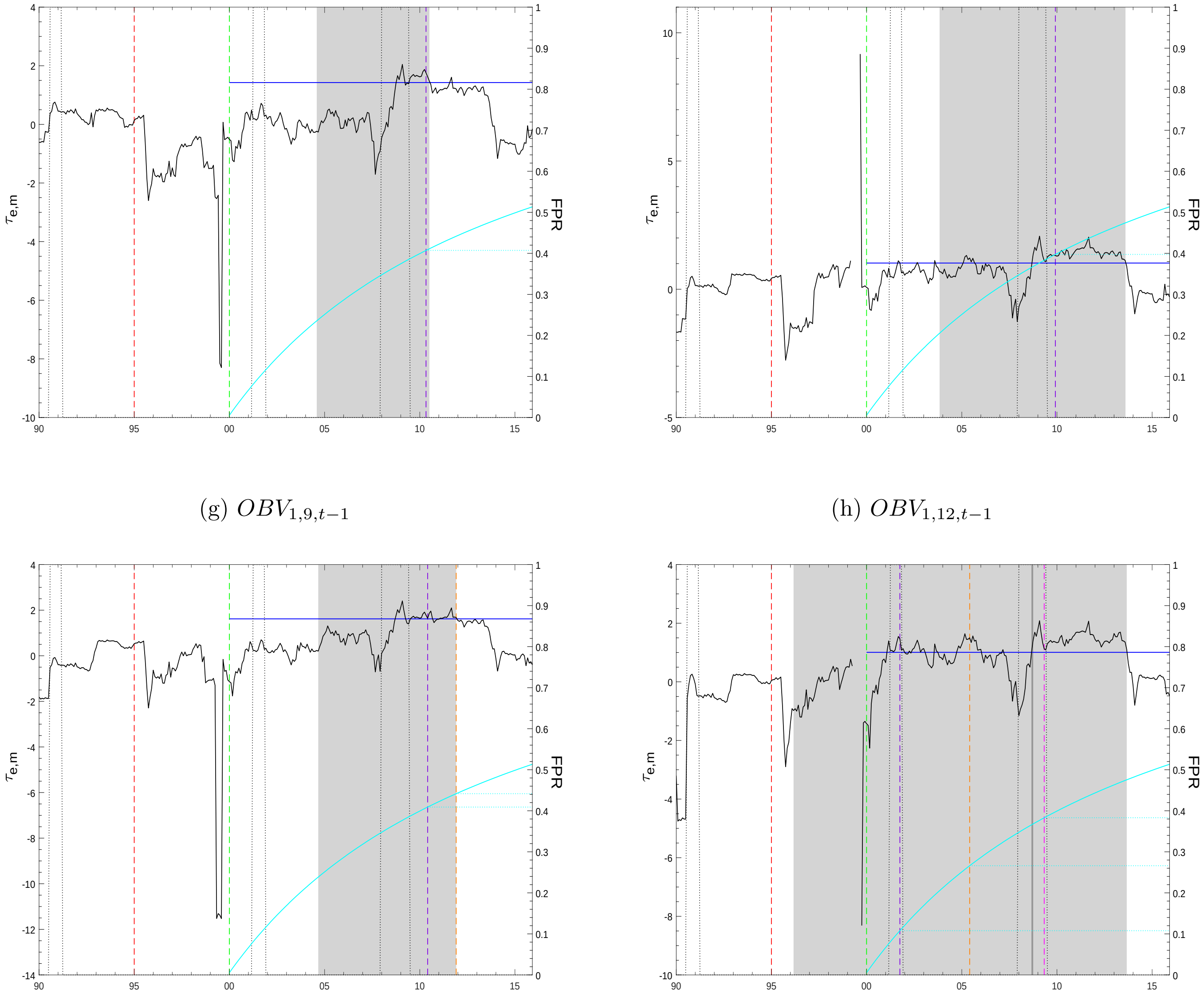
(c) $MAI_{2,9,t-1}$

(d) $MAI_{2,12,t-1}$

(e) $MOI_{9,t-1}$

(f) $MOI_{12,t-1}$

Figure S13. *SEQ* procedure, $m = 60$: —($\tau_{e,m}$), —($cv_{0.10}$), - - -(T^*), - - -(T^*+m), - - -(first rejection), - - -(second rejection), - - -(third rejection), ▨ (weak set of dates), ■ (strong set of dates), —(false positive rate), ·····(NBER indicator)



(g) $OBV_{1,9,t-1}$

(h) $OBV_{1,12,t-1}$

(i) $OBV_{2,9,t-1}$

(j) $OBV_{2,12,t-1}$

Figure S13 Continued. *SEQ* procedure, $m = 60$: —($\tau_{e,m}$), —($cv_{0.10}$), - - -(T^*), - - -(T^*+m), - - -(first rejection), - - -(second rejection), - - -(third rejection), ▨ (weak set of dates), ■ (strong set of dates), —(false positive rate), ·····(NBER indicator)