Monitoring Forecasting Performance*

Allan Timmermann† Yinchu Zhu‡

Abstract

We develop conditions under which the expected predictive accuracy of a set of competing forecasting models can be ranked either unconditionally, based on their average performance, or conditionally, based on a set of time-varying monitoring instruments. We characterize the properties that monitoring instruments must possess and show that these reflect both the accuracy of the predictors used by the competing nested or non-nested forecasting models and the strength of the monitoring instruments. We derive finite-sample bounds on forecasting performance that account for estimation error both in the underlying forecasting models and in monitoring regressions used to compute the expected loss of the forecasts. We quantify the expected gains from monitoring forecasting performance for a decision maker that, at each point in time, chooses between competing forecasts conditional on information in the monitoring instruments. Using Monte Carlo simulations and empirical applications to inflation forecasting and predictability of stock market returns, we demonstrate the gains from monitoring forecasting performance.

*We thank Frank Diebold, Hashem Pesaran and seminar participants at USC, the Tinbergen Institute at Erasmus University, Rotterdam, the 2016 Duke conference on New Developments in Measuring & Forecasting Financial Volatility, the 2017 Sofie conference in New York City, and the 2017 ISF conference in Cairns for comments on the paper.

†Rady School of Management, University of California, San Diego, 9500 Gilman Dr, La Jolla, CA 92093, U.S.A.; atimmermann@ucsd.edu

‡Lundquist College of Business, University of Oregon, 1208 University St, Eugene, OR 97403, U.S.A.; yzhu6@uoregon.edu

1 Introduction

Large international organizations such as the IMF and the World bank as well as central banks all over the world routinely generate economic forecasts that are widely followed and play a key role in these organizations' decision process.[1] Monitoring the performance of these forecasts in "real time" is crucial as evidence that forecasting performance is deteriorating or particularly poor in certain economic states could be used either for gauging the likely accuracy of a forecast at a given point in time or to improve on an existing forecasting approach and switch to a better one. Indications that a forecast is likely to be surrounded by higher-than-normal uncertainty should reduce decision makers' dependence on that forecast. In this regard, monitoring forecasting performance is similar to the task of an asset manager who is evaluating portfolio risk at regular points in time.

Given the presence of an often wide range of alternative model specifications and different methods for implementing a particular model, it is not surprising that a large academic literature has been devoted to developing ways for comparing predictive accuracy.[2] Such comparisons typically consider whether one forecast "on average" (or unconditionally) is more accurate than other, competing, forecasts. However, it is possible for forecasts to be poor on average, yet still be relatively accurate in some states of the world. Provided that these states can be ex-ante identified by means of a set of time-varying monitoring instruments, a forecast that is poor on average could be the preferred forecast at a given point in time, conditional on information contained in the monitoring instruments.

The existence of monitoring instruments that contain information either on a particular forecast's *absolute* performance, measured relative to the outcome, or its *relative* performance measured against competing forecasts is important given the widespread empirical evidence of model instability in macroeconomics and finance.[3] This evidence suggests that it is rare to find a single forecasting model that uniformly

[1] In fact, the IMF regularly reviews the performance of its World Economic Outlook forecasts, comparing the accuracy of their forecasts to those of other organizations such as Consensus Economics. See Timmermann (2007) for such a comparison.

[2] See Granger and Newbold (1977), Chong and Hendry (1986), Diebold and Mariano (1995), and West (1996) for early work on comparison of different models' predictive accuracy. Clark and McCracken (2013) provide a review of recent work in the literature.

[3] See, e.g., Giacomini and Rossi (2009), Pesaran et al. (2006), Pettenuzzo and Timmermann (2015), Rossi (2013), Rossi and Sekhposyan (2015), and Stock and Watson (1996).

dominates other forecasts through time.[4] More broadly, time variation in the expected performance of different forecasts can make it attractive to select the forecast that, at a given point in time and conditional on a set of monitoring instruments, is expected to produce the most accurate forecast.

To understand the importance of the strength of the predictor(s) used by a particular forecast, we follow the literature and consider the case where the forecasts are generated by a set of linear forecasting models whose parameters get updated recursively through time. This setup allows us to rank the expected performance of different models and quantify the effect of parameter estimation error. Accounting for estimation error poses technical challenges to the analysis but turns out to be crucial for understanding the models' forecasting performance. We first consider the case in which one model uses a strictly stronger predictor than its competitor. Under the assumption that the strong predictor's signal is more powerful than local to zero, we show that bounds can be established on the two models' mean squared error (MSE) loss difference. Importantly, our uniform bounds hold in finite samples and account for estimation error. As expected, the bounds on the expected reduction in MSE from using the best model depends on the strength of the predictor.

The idea of conducting forecast comparisons conditional on time-varying instruments was introduced to economics by Diebold and Mariano (1995) and Giacomini and White (2006) but has not been widely pursued, nor have conditions been established under which monitoring instruments with predictive power over loss differentials will exist or the possible gains from monitoring forecasting performance when two or more forecasts are available. We characterize in this paper the properties that monitoring instruments must possess in order to contain valuable information about the competing models' forecasting performance. We first consider the case with nonnested forecasting models in which each forecasting model uses at least one predictor that is excluded by the other model. For this case a monitoring instrument can be used to track time variation in the models' relative squared error forecasting performance if the instrument is sufficiently strongly correlated with the cross-product of the residual and predictors included by one model but excluded by others. Both the strength of the predictor and of the monitoring instrument turn out to matter for our

[4] Stock and Watson (1996) find that model instability impacts a majority of a large range of macroeconomic variables. Rossi and Sekhposyan (2013) arrive at similar conclusions and provide a thorough review of the literature on how model instability affects forecasting performance.

ability to rank different models' conditional forecasting performance, and we present finite-sample bounds on the expected gain from using a switching rule that chooses the best model conditional on the monitoring instruments, relative to always using a particular model. These bounds account for estimation error both in the underlying forecasting model and in the monitoring (switching) regression.

We next consider the nested case using a setting with a "small" (benchmark) model and a "large" competing model that nests the benchmark as a special case. If the additional predictor included in the large model is either uninformative or very strong, one of the models can be expected to always produce better forecasts. Alternatively, when the additional predictor contains "weak" information (i.e., its coefficients in the forecasting model is at most local-to-zero) the identity of the model with the best expected performance may vary over time in a way that is correlated with information in the monitoring instruments.[5]

As an empirical illustration of the approach developed in this paper, consider the predictive accuracy of the Greenbook inflation forecasts, published by the US Federal Reserve, relative to that of inflation forecasts from the Survey of Professional Forecasters (SPF).[6] The left window in Figure 1 shows (in blue) the squared error differential between the Greenbook and SPF forecasts using one-quarter-ahead forecasts of the growth in the GDP price deflator. Negative values indicate that the Greenbook forecast was more accurate (produced a smaller squared forecast error) in a particular quarter than the corresponding SPF forecast. The figure shows that Greenbook forecasts were generally more accurate than the SPF forecasts in the early part of the sample, while the converse holds for most of the last part of the sample. We also show (in red) forecasts of squared error differentials generated by recursive regressions of squared error differentials on an intercept and the lagged value of the macroeconomic uncertainty factor constructed by Jurado et al. (2015).[7] Negative forecasts of the

[5]Under model misspecification arising from time-varying parameters or due to the use of the wrong functional form of the forecasting model, monitoring instruments can contain valuable information if they are correlated with the model specification error.

[6]This is clearly a case of non-nested forecasts. Comparisons of the predictive accuracy of the Federal reserve to that of private sector forecasters have been the subject of previous research such as Romer and Romer (2000) and Capistrán (2008) and have important policy implications. The earlier studies were concerned with comparing the average performance of these forecasts; our approach instead focuses on comparing their conditional forecasting performance so as to track possible shifts in the forecasts' relative accuracy.

[7]The correlation between this instrument and the squared error loss difference for the Greenbook versus the SPF forecasts is -0.18, suggesting that the Greenbook forecasts tend to do relatively better

squared error differential indicate that the expected loss for the Greenbook forecasts are smallest while positive values indicate that the SPF forecasts are expected to be more accurate. These forecasts are then used in a simple switching rule that chooses the Greenbook forecast if its conditionally expected loss is smaller than that of the SPF forecast, otherwise goes with the SPF forecast.[8] The Greenbook forecasts are selected in most quarters up to 2000, whereas the SPF forecasts are selected most of the time thereafter.

The right window in Figure 1 tracks the cumulative loss difference associated with different forecasting strategies. The blue line shows the cumulative difference in the squared forecast errors from the Greenbook minus their SPF counterpart. Whenever this line is negative and declining, the Greenbook forecasts generate a lower squared error loss than the SPF forecasts. Conversely, positive and increasing values of the line suggest that the SPF forecasts are more accurate than the Greenbook forecasts at that point in time. Greenbook forecasts were better during most of the 1980s, the two sets of forecasts performed broadly similarly during the 1990s, while the SPF forecasts have been better since 2000, with the notable exception of the financial crisis of 2008/2009. The other lines track the cumulative squared error performance of forecasts from the switching rule measured relative to the Greenbook forecasts (in green) or relative to the SPF forecasts (in red). The conditional switching rule performs as well as the Greenbook forecasts up to 2003, but outperforms after this period as a result of successfully identifying the improved performance of the SPF forecasts in the latter half of the sample. Measured against the SPF forecasts, the performance of the conditional switching rule is even better at the end of the sample.

This application shows that our method for monitoring forecasting performance can be beneficial for decision makers such as central banks in "benchmarking" their forecasts against forecasts from other sources to see if there is evidence in real time that the quality of their forecasts is lagging behind competing forecasts. While it may not be feasible for such decision makers to adopt another agency's forecasts, evidence of inferior performance would suggest the need to improve on their own forecasting methods.

A key difference between the conditional forecast evaluation approach considered here and conventional unconditional tests of forecasting performance, highlighted by

than the SPF forecasts during times with high levels of macroeconomic uncertainty.

[8]This regression uses a rolling window of 10 years, i.e., 40 quarterly observations.

Figure 1, is that the former allows us to identify the economic sources (states) of time variation in a model's predictive ability. This type of dynamic attribution analysis can be helpful in two regards. First, at a minimum, a rational decision maker will use a forecast with less confidence in economic environments where the forecasts are expected to perform poorly compared to states in which the forecast is expected to perform well. Second, information on a forecast's conditional performance might be used to suggest ways in which the underlying forecasting method could be improved. For example, knowing that a forecast is poor relative to some other benchmark when macroeconomic uncertainty is high is a lot more informative than simply knowing that a forecast is poor on average.

Our analysis builds on the seminal paper by Giacomini and White (2006) which formalizes the notion of conditional forecast evaluation in the context of regression-based tests that can capture non-zero correlations between a set of monitoring instruments and the predictive accuracy of a particular forecasting model measured relative to some benchmark. However, these authors do not develop conditions under which such monitoring instruments exist, let alone when they should not simply be added to the underlying forecasting models. Nor do they characterize the finite-sample expected gains from a conditional decision rule that, at each point in time, chooses the model with the best expected forecasting performance. Their analysis also does not relate time variation in predictive accuracy to parameter estimation error and the notion of "weak" predictors, which is central to our analysis.

Finally, our analysis suggests alternative monitoring instruments that can be expected to track forecasting performance in real time. One approach is to simply use the lagged loss differential or the lagged loss of the individual forecasting models. These instruments can capture highly persistent sources of forecast errors such as parameter estimation error. Another monitoring instrument is the lagged difference between the squared forecasts. Provided that the differences in the squared values of the predictors are persistent, lags of the loss differential can be helpful for computing which model is expected to have the best future performance. Neither of these instruments requires that the predictor variables underlying the forecasting models are observed and can be used in situations where we only observe the outcome and a sequence of forecasts.

The outline of the paper is as follows. Section 2 introduces the setup for the analysis, including the conditional testing methodology. Sections 3, 4, and 5 contain

our formal theoretical analysis which establishes how competing forecasting models, as well as a conditional decision rule, can be ranked. We first perform the analysis for the general non-nested case (Section 3) and then cover the case with nested forecasting models (Section 4). Section 5 studies the case without parameter estimation errors and extends the analysis to cover highly persistent (near unit root) predictors. Section 6 reports the outcome of a set of Monte Carlo simulations that we use to illustrate the theoretical analysis, while Section 7 presents an empirical analysis of predictability of US inflation and stock market returns. Section 8 concludes.

2 Conditional Tests of Equal Predictive Accuracy

Conventional tests of equal predictive accuracy inspect whether alternative forecasts are equally accurate, on average, given a sample of forecasts and outcomes. Suppose, however, that one forecast is expected to be more accurate than another forecast conditional on some state variable (monitoring instrument) taking certain values, while the ranking of the two forecasts' expected performance is reversed for other values of the state variable. In this case, the strategy of always using the forecast that is most accurate "on average" could be dominated by a decision rule that selects the forecast that, at each point in time, has the smallest expected loss.

This point highlights the value from implementing conditional tests of equal predictive accuracy though it does not establish the conditions under which such state variables exist or the conditions under which such tests might work. In this section we introduce more formally the forecast environment and discuss ways to test for equal conditional and unconditional forecasting performance. Lastly, we introduce a switching rule that exploits the conditional information in the monitoring instruments to choose between forecasting models.

2.1 Out-of-Sample Tests of Forecasting Performance

Pairwise comparisons of predictive accuracy are now routinely carried out in macroeconomic and financial studies.[9] Typically it is assumed that forecasts are generated from a set of underlying linear models whose parameters are updated recursively as new information arrives. In the same spirit, it is common practice to compute

[9]For a recent review of the literature, see Clark and McCracken (2013).

tests of relative predictive accuracy by studying the models' out-of-sample forecasting performance. Out-of-sample tests assume that part of the sample is reserved for initial estimation of model parameters and that these parameters get updated recursively (using a rolling or an expanding estimation window) as new information arrives. Forecasting performance is evaluated using these recursively generated forecasts, thus avoiding that the same sample of data is used to estimate model parameters and evaluate the resulting forecasts. Out-of-sample performance is studied, first, to address whether economically useful forecasts could have been generated in real time and, second, to guard against data mining or overfitting biases.[10]

In common with much of the literature, we shall focus on univariate forecasting problems. Specifically, let $\hat{y}_{1,t+1|t}$ and $\hat{y}_{2,t+1|t}$ be a set of one-step-ahead forecasts of the outcome y_{t+1} generated using information known at time t.[11] Following Diebold and Mariano (1995), we can evaluate the accuracy of the forecasts using a loss function $L(\hat{y}_{t+1|t}, y_{t+1})$, where $\hat{y}_{t+1|t} \in \{\hat{y}_{1,t+1|t}, \hat{y}_{2,t+1|t}\}$. By far the most common loss function is squared error loss

$$L(\hat{y}_{t+1|t}, y_{t+1}) = (y_{t+1} - \hat{y}_{t+1|t})^2. \tag{1}$$

Under squared error loss, the loss differential between two forecasts, $\Delta L_{t+1} \equiv L(\hat{y}_{1,t+1|t}, y_{t+1}) - L(\hat{y}_{2,t+1|t}, y_{t+1})$, takes the form

$$\Delta L_{t+1} = e_{1,t+1}^2 - e_{2,t+1}^2, \tag{2}$$

where $e_{j,t+1} = y_{t+1} - \hat{y}_{j,t+1|t}$ for $j = 1, 2$ are the individual forecast errors. Negative values of ΔL_{t+1} in (2) show that the first forecast produced a smaller squared forecast error than the second forecast in period $t + 1$.

2.2 Tests of Equal Unconditional Forecasting Performance

Following Diebold and Mariano (1995), we can test if the two forecasts have the same unconditional expected ("average") loss through the null hypothesis

$$H_0 : \mathbb{E}[\Delta L_{t+1}] = 0, \tag{3}$$

where $\mathbb{E}[.]$ is the expectation operator.

[10] See Hansen and Timmermann (2015a) for further discussion of this point.

[11] Note that our setup can easily be generalized to allow for a multi-period forecast horizon.

The null in (3) has been tested extensively in empirical studies in economics and finance. The simplest way of testing if one of the forecasts dominates the other on average is to test if the mean of the sequence of loss differentials ΔL_{t+1} is non-zero through a Diebold and Mariano (1995) regression

$$\Delta L_{t+1} = \theta_0 + u_{t+1}. \tag{4}$$

A t-test can be used to test if $\theta_0 \neq 0$, in which case the average MSE performance of the two forecasts is significantly different from zero. Provided that ΔL_{t+1} is covariance stationary, such $t-$tests will follow a standard distribution as discussed in Diebold (2015). The effect of recursive updating of parameter estimates when the forecasts are generated by linear forecasting models has been considered by West (1996) (for non-nested models) and by Clark and McCracken (2001), McCracken (2007) and Hansen and Timmermann (2015b) (for nested models).

2.3 Tests of Equal Conditional Forecasting Performance

Even if one forecast is worse *on average* than another forecast, it might perform better in certain states of the world. This suggests using a conditional test of predictive accuracy that conditions on observable information when evaluating competing forecasts' relative accuracy. Giacomini and White (2006) (GW, henceforth) propose a method for doing this. Let $\mathcal{G}_t$ denote the information set consisting of variables used to assess the two forecasts' predictive accuracy. GW extend unconditional tests implied by the null in (3) to conditional tests based on the null

$$H_0 : \mathbb{E}[\Delta L_{t+1}|\mathcal{G}_t] = 0. \tag{5}$$

Note that the null in (3) could be true even if (5) is false: Two forecasts may generate the same average loss even though information in $\mathcal{G}_t$ can be used to predict when one forecast performs better than the other. Conversely, if the null in (5) holds for all elements in $\mathcal{G}_t$, then (3) follows trivially, assuming that $\mathcal{G}_t$ includes a constant.

Following GW, we can turn the null in (5) into a test statistic by using a set of monitoring instruments ("testers") $Z_t \in \mathcal{G}_t$, so that we test $H_0 : \mathbb{E}[\Delta L_{t+1}|Z_t] = 0$ where $Z_t \in \mathbb{R}^{d_Z} \subseteq \mathcal{G}_t$ is a sequence of conditioning variables which typically includes a constant and d_Z is the dimension of Z_t.

Under the null in (5), Z_t should be orthogonal to the loss differential ΔL_{t+1}. This can be tested using a linear regression:

$$\Delta L_{t+1} = (\theta_0, \theta_1) \begin{pmatrix} 1 \\ z_{1t} \end{pmatrix} + u_{t+1} \equiv \theta' z_t + u_{t+1}, \tag{6}$$

where $z_t = (1,\ z_{1t})'$, $\mathbb{E}[u_{t+1} z_{1t}] = 0$, and $z_{1t} \in Z_t$.[12]

Under the null of equal conditional predictive accuracy, $\theta_0 = 0$ and $\theta_1 = 0$ in (6). Non-zero values of θ_1 suggest that the monitoring instruments, z_{1t}, can help forecast differences in predictive accuracy across the two forecasts.

Under the moment and mixing conditions in Theorem 1 of Giacomini and White (2006) the null hypothesis in (5) can be tested on a sample $\{y_{t+1}, \hat{y}_{1,t+1|t}, \hat{y}_{2,t+1|t}, z_{1t}\}_{t=T-p}^{T-1}$ by computing the GW test statistic

$$\begin{aligned} J_T &= \left[p^{-1/2} \sum_{t=p}^{T-1} \Delta L_{t+1} Z_t \right]' \hat{\Omega}_T^{-1} \left[p^{-1/2} \sum_{t=p}^{T-1} \Delta L_{t+1} Z_t \right] \\ &\xrightarrow{d} \chi^2 (d_Z), \end{aligned} \tag{7}$$

where $\hat{\Omega}_T = p^{-1} \sum_{t=p}^{T-1} (\Delta L_{t+1} Z_t - \overline{\mu}_T)(\Delta L_{t+1} Z_t - \overline{\mu}_T)'$ is a consistent estimate of the variance of $\Delta L_{t+1} Z_t$, and $\overline{\mu}_T = p^{-1} \sum_{t=p}^{T-1} \Delta L_{t+1} Z_t$ is the mean of the product of the loss differential and the d_Z monitoring instrument Z_t.

2.4 Expected Gains from Monitoring

Suppose the test in (7) rejects that $\mathbb{E}(\Delta L_{t+1} | Z_t) = 0$ for some monitoring instrument, Z_t, suggesting that Z_t can be used to predict the mean of ΔL_{t+1} and tell when one forecast is likely to perform better than the other. Using the monitoring regression in (6), we can compute the expected future loss $\mathbb{E}(\Delta L_{t+1} | Z_t)$ by $\theta' Z_t$. This expectation can be the basis for choosing which forecast to use for period $t+1$. Following GW, we consider a simple switching rule that chooses forecast 1 if $\mathbb{E}(\Delta L_{t+1} \mid Z_t) \leq 0$, otherwise chooses forecast 2:

$$\hat{y}_{SW,t+1|t} = \hat{y}_{1,t+1|t} \{\mathbb{E}[\Delta L_{t+1} | z_t] \leq 0\} + \hat{y}_{2,t+1|t} \mathbf{1} \{\mathbb{E}[\Delta L_{t+1} | z_t] > 0\}, \tag{8}$$

[12] Non-linearities can easily be incorporated by using transformations of z_{1t}.

where $\mathbf{1}\{\mathbb{E}[\Delta L_{t+1}|z_t] > 0\}$ is an indicator variable that equals one if the first forecasting model has the highest expected loss conditional on $Z_t = z_t$, otherwise is zero.

To establish if the instrument Z_t can be useful in tracking the relative performance of the two forecasts, note that if $\mathbb{E}[\Delta L_{t+1}|z_t] < 0$ with positive probability–so the conditionally expected loss of the second forecast exceeds that of the first forecast for some value of z_t–then $\mathbb{E}[\Delta L_{t+1}] < \mathbb{E}[\Delta L_{t+1}\mathbf{1}\{\mathbb{E}(\Delta L_{t+1}|z_t) > 0\}]$.[13] Moreover, the amount by which the switching rule (8) is expected to outperform forecast 1, conditional on forecast 2 having the lowest expected loss, is $\mathbb{E}[\Delta L_{t+1}\mathbf{1}\{\mathbb{E}[\Delta L_{t+1}|z_t] > 0\}]$. Equivalently, the expected gain from using the switching rule (8), conditional on forecast 1 having the lowest expected loss, is $\mathbb{E}[-\Delta L_{t+1}\mathbf{1}\{\mathbb{E}[\Delta L_{t+1}|z_t] \leq 0\}]$ and the switching rule is expected to outperform forecast 1 unconditionally provided that $\mathbb{E}(\Delta L_{t+1} \mid Z_t) > 0$ occurs with strictly positive probability.

These arguments show that there are gains expected from forecast monitoring relative to always using forecasts from a particular model provided that neither of the underlying forecasting models is too dominant since $\mathbb{E}[\Delta L_{t+1}|z_t]$ is required to switch sign for different values of z_t. These conclusions only assume the existence of $\mathbb{E}[|\Delta L_{t+1}|]$ and so are established under very weak conditions. However, they do not account for estimation error, nor do they quantify the potential benefits from monitoring forecasting performance and, at each point in time, selecting the forecast with the lowest conditionally expected loss. This is the topic of the next two sections.

3 Comparing Forecasts from Non-nested Models

This section studies pair-wise comparisons of forecasts generated by non-nested models, both of which only partially captures the information in the data generating process for y_{t+1}. The case with nested models is covered in the next section. We derive conditions under which approximate finite sample bounds can be established on the relative performance of non-nested forecasting models in the presence of estimation error and also characterize conditions under which a switching rule can be

[13] To see this, notice that if $\mathbb{E}[\Delta L_{t+1}|z_t] < 0$ with positive probability, then

$$\begin{aligned}\mathbb{E}[\Delta L_{t+1}] - \mathbb{E}[\Delta L_{t+1}\mathbf{1}\{\mathbb{E}[\Delta L_{t+1}|z_t] > 0\}] &= \mathbb{E}[\Delta L_{t+1}\mathbf{1}\{\mathbb{E}(\Delta L_{t+1}|z_t) \leq 0\}] \\ &= \mathbb{E}[\{\mathbb{E}[\Delta L_{t+1}\mathbf{1}\{\mathbb{E}(\Delta L_{t+1}|z_t) \leq 0\}|z_t]\}] < 0.\end{aligned}$$

expected to produce better forecasts than either model.

3.1 Pairwise Comparisons

Consider the data generating process (DGP)

$$y_{t+1} = \beta_1 x_{1,t} + \beta_2 x_{2,t} + \varepsilon_{t+1}, \tag{9}$$

where x_{1t} and x_{2t} are a set of predictor variables that are known at time t. To capture the case with non-nested models, we assume that model 1 takes the form $y_{t+1} = \beta_1 x_{1,t} + \varepsilon_{1t+1}$, while model 2 takes the form $y_{t+1} = \beta_2 x_{2,t} + \varepsilon_{2t+1}$. For simplicity, we assume that x_{1t} and x_{2t} are univariate processes.[14]

Our analysis assumes that we observe a sample of T data points $\{(x_{j,t}, y_t)\}_{t=1}^{T}$. To allow for mild time variation in the parameters of the prediction models, in line with standard practice in the literature on out-of-sample forecasting we assume that the parameters of the forecasting models ($\hat{\beta}_{j,n,t}$) are estimated using a rolling window of the most recent n observations

$$\hat{\beta}_{j,n,t} = \left(\sum_{s=t-n}^{t-1} x_{j,s}^2 \right)^{-1} \left(\sum_{s=t-n}^{t-1} x_{j,s} y_{,s+1} \right), \ j = 1, 2. \tag{10}$$

The resulting forecasts are generated as $\hat{y}_{j,t+1|t} = \hat{\beta}_{j,n,t} x_{j,t}$ for $j = 1, 2$. Under this setup, the sample size, T, is split into a rolling window of length n used to estimate $\hat{\beta}_{j,n,t}$, the parameters of the jth forecasting model, and an evaluation sample containing the remaining p observations so that $T = n + p$. Both n and p can be functions of T and may or may not tend to infinity but, for simplicity, we write n and p instead of n_T and p_T.[15] The tests that evaluate the performance of the two models are based on the p forecasts $\{(\hat{y}_{1,t+1|t}, \hat{y}_{2,t+1|t})\}_{t=n}^{n+p-1}$.

Using (9), the squared error loss differential ΔL_{t+1} becomes

$$\begin{aligned} \Delta L_{t+1} &= (y_{t+1} - \hat{y}_{1,t+1|t})^2 - (y_{t+1} - \hat{y}_{2,t+1|t})^2 \\ &= 2\varepsilon_{t+1} \left(\beta_2 x_{2,t} - \beta_1 x_{1,t} - \delta_{1,n,t} x_{1,t} + \delta_{2,n,t} x_{2,t} \right) \end{aligned}$$

[14] Note that we do not rule out that β_1 or β_2 equal zero.

[15] Giacomini and White (2006) assume that the length of the estimation window, n, is fixed. However, Timmermann and Zhu (2016) generalize this setup to allow for an expanding estimation window.

$$+ \left(\beta_2 x_{2,t} - \beta_1 x_{1,t} - \delta_{1,n,t} x_{1,t} + \delta_{2,n,t} x_{2,t}\right)\left(\beta_1 x_{1,t} + \beta_2 x_{2,t} - \delta_{1,n,t} x_{1,t} - \delta_{2,n,t} x_{2,t}\right), \tag{11}$$

where $\delta_{j,n,t} = \hat{\beta}_{j,n,t} - \beta_j$ denotes the estimation error for model j.

3.2 Determinants of Average Forecasting Performance

To ensure that our analysis of forecasting performance allows for a broad set of time-series dependencies, we adopt the β-mixing condition formulated as in Section 2.1 of Chen et al. (2016) adapted to an array setting similarly to Andrews (1988):

Definition 1. The array $\{W_{T,t}\}_{t=-\infty}^{\infty}$ is said to be β-mixing with coefficient $\beta_{mix}(\cdot)$ if

$$\beta_{mix}(t) = \sup_{-\infty<i<\infty, T\geq 1} \mathbb{E}\left(\sup_{B\in\mathcal{F}_{i+t,\infty}^{T}} \left|\mathbb{P}\left(B|\mathcal{F}_{-\infty,i}^{T}\right) - \mathbb{P}(B)\right|\right) \to 0 \; as \; t \to \infty,$$

where $\mathcal{F}_{-\infty,i}^{T} = \sigma(\cdots, W_{T,i-1}, W_{T,i})$ and $\mathcal{F}_{i+t,\infty}^{T} = \sigma(W_{T,i+t}, W_{T,i+t+1}, \cdots)$.

The array structure in this definition is general enough to allow for many types of nonstationary data and provides a convenient way of analyzing data generating processes indexed by the sample size. For example, we can allow the signal strength of the predictor, x, and monitoring instrument, z, to decay to zero as T gets large.

Characterizing the expectation of ΔL_{t+1} in a way that is relevant for testing purposes is technically challenging. Standard asymptotic results, such as the classical central limit theorem (CLT), require n to tend to infinity in order to approximate the behavior of $\delta_{j,n,t}$. For example, for a fixed t, one can use a CLT to show that $\delta_{j,n,t} = O_P(n^{-1/2})$ and thus $\Delta L_{t+1} = \beta_2^2 x_{2,t}^2 - \beta_1^2 x_{1,t}^2 + 2(\beta_2 x_{2,t} - \beta_1 x_{1,t})\varepsilon_{t+1} + O_P(n^{-1/2})$. For $\beta = cn^{-\alpha_x}$ and large n, with $\alpha_x < 1/4$ the leading term of $\mathbb{E}\Delta L_{t+1}$ is $\beta_2^2 \mathbb{E}x_{2,t}^2 - \beta_1^2 \mathbb{E}x_{1,t}^2$ since $\mathbb{E}\left[x_{j,t}\varepsilon_{t+1}\right] = 0$. However, for $\alpha_x > 1/4$, it is unclear how $\mathbb{E}\Delta L_{t+1}$ behaves and the outcome of a GW test on ΔL_{t+1} is not obvious.

Moreover, the potential lack of uniformity in the $O_P(n^{-1/2})$ terms presents challenges. Since the $O_P(n^{-1/2})$ term might not be uniform in t, one cannot conclude that performing a GW test on ΔL_{t+1} is asymptotically equivalent to implementing the test using $\beta_2^2 x_{2,t}^2 - \beta_1^2 x_{1,t}^2 + 2(\beta_2 x_{2,t} - \beta_1 x_{1,t})\varepsilon_{t+1}$, even if $\alpha_x < 1/4$. The reason is that, for an array of random variables, each of which converges to zero in probability,

the average of these random variables need not converge to zero in probability.[16]

To overcome these difficulties, we derive bounds that are valid uniformly across t, hold in finite samples, and thus do not require an asymptotic framework. To this end, we make use of the following list of assumptions:

Assumption 1. *The following hold for $j \in \{1, 2\}$:*

(i) There exist constants $r > 8$ and $D > 0$ such that $\mathbb{E}|x_{j,t}|^r$ and $\mathbb{E}|\varepsilon_{t+1}|^r$ are bounded above by D. Moreover, $\mathbb{E}x_{j,t} = \mathbb{E}x_{1,t}x_{2,t} = \mathbb{E}x_{j,t}\varepsilon_{t+1} = 0$.

(ii) $\{x_{j,t}, \varepsilon_{t+1}\}_{t=-\infty}^{\infty}$ is a β-mixing array with coefficient $\beta_{mix}(\cdot)$ such that $\forall t > 0$, $\beta_{mix}(t) \leq b\exp(-t^c)$, for constants $b, c > 0$. Moreover, for some constants $Q_1, Q_2 > 0$, $\mathbb{E}(k^{-1/2}\sum_{s=t-k}^{t-1} x_{1,s}(x_{2,s}\beta_2 + \varepsilon_{s+1}))^2$, $\mathbb{E}(k^{-1/2}\sum_{s=t-k}^{t-1} x_{2,s}(x_{1,s}\beta_1 + \varepsilon_{s+1}))^2$, $\mathbb{E}x_{1,t}^2$ and $\mathbb{E}x_{2,t}^2$ lie in $[Q_1, Q_2]$ for all k, t.

(iii) There exist constants $\alpha_{x,j} \in [0, \infty]$, $c_{\beta,j} > 0$ such that $\beta_j = c_{\beta,j}n^{-\alpha_{x,j}}$, where

(a) $\alpha_{x,2} < \alpha_{x,1}$,

(b) $\alpha_{x,2} < 1/2$.

(iv) $n/T > \kappa$ for some constant $\kappa > 0$.

Assumption 1(i) imposes relatively weak moment conditions on $x_{j,t}$ and ε_{t+1} and holds for many processes, including many GARCH specifications. Notice that we do not require exponential-type tails, which are routinely imposed in papers that handle uniformly valid bounds; see e.g., Fan et al. (2011) and Bonhomme and Manresa (2015). The mixing condition in Assumption 1(ii) ensures weak dependence in the data and is commonly used in the literature. Importantly, we do not impose stationarity and allow for heteroskedasticity. Assumption 1(iii) characterizes the strength of the predictors through the order of magnitude of their coefficients in the forecasting model, $\beta_j = c_{\beta,j}n^{-\alpha_{x,j}}$. The smaller the value $\alpha_{x,j}$, the stronger the predictor, with $\alpha_{x,j} = 0$ representing the conventional case with a very strong predictor whose presence can be detected with certainty as the sample size increases, while $\alpha_{x,j} = 1/2$ represents the local-to-zero case with a weaker predictor whose importance is much

[16] For example, consider the array $a_{n,t} = t^4/n$: for a fixed t, $\lim_{n\to\infty} a_{n,t} = 0$ but the average $n^{-1}\sum_{t=1}^{n} a_{n,t} \to \infty$.

harder to detect. Without loss of generality, we assume that the predictor in the second model is stronger than the predictor in the first model ($\alpha_{x,2} < \alpha_{x,1}$) and we further assume that the dominant predictor is stronger than local-to-zero ($\alpha_{x,2} < 1/2$). The case with local-to-zero predictors ($\alpha_{x,2} = 1/2$) has been the subject of many studies including, most recently, Hirano and Wright (2017). Studying sequences of parameter values whose magnitude declines as the sample size grows bigger ensures that parameter uncertainty is preserved asymptotically. In contrast, with a fixed alternative ($\alpha_{x,2} = 0$), uncertainty about the parameter estimates disappears asymptotically.

Finally, assumption 1(iv) requires that the estimation window, n, grows at (at least) the same rate as the sample size, T.

The following result allows us to study the finite-sample properties of the expected squared error performance of the two models:

Proposition 1. *Consider the DGP*

$$y_{t+1} = \beta_1 x_{1,t} + \beta_2 x_{2,t} + \varepsilon_{t+1},$$

Moreover, assume that the parameters of the forecasting models, $\hat{\beta}_{j,n,t}$, are estimated recursively using n observations. Then, under squared error loss and Assumption 1,

(1) there exist constants $G_1, G_2 > 0$ and an array of random variables $\{\Delta L_{t+1,*}\}_{t=n}^{T-1}$ such that for $T \geq G_1$,

$$\mathbb{P}\left(\bigcap_{t=n}^{T-1} \{\Delta L_{t+1,*} = \Delta L_{t+1}\}\right) \geq 1 - G_2 T^{\max\{1-r/8,\ 1+(\alpha_{x,2}-1/2)(2+r/4)\}}.$$

(2) there exist constants $G_3, G_4 > 0$ such that, for $n \leq t \leq T-1$,

$$G_3 T^{-2\alpha_{x,2}} \leq \mathbb{E}\Delta L_{t+1,*} \leq G_4 T^{-2\alpha_{x,2}}.$$

Part 1 of Proposition 1 establishes a coupling result that allows us to study the behavior of ΔL_{t+1}. Since ΔL_{t+1} might not have bounded moments for finite n, we consider $\{\Delta L_{t+1,*}\}$, which coincides with $\{\Delta L_{t+1}\}$ with high probability. Specifically, if $T^{\max\{1-r/8,\ 1+(\alpha_{x,2}-1/2)(2+r/4)\}}$ vanishes, i.e., $r > \max\{8,\ 16\alpha_{x,2}/(1-2\alpha_{x,2})\}$, tests computed based on $\{\Delta L_{t+1}\}_{t=n+m}^{n+m+p-1}$ have the same asymptotic properties as those computed using $\{\Delta L_{t+1,*}\}_{t=n+m}^{n+m+p-1}$.

Our result provides finite-sample properties of $\mathbb{E}\Delta L_{t+1,*}$ for each t, n and T without requiring T or n to tend to infinity. Further, notice that ΔL_{t+1} is not necessarily stationary. For example, if n grows with T or (x_t, ε_t) is not stationary even for fixed n, then ΔL_{t+1} will not, in general, be stationary.

Proposition 1 shows that Model 2 is expected to generate a smaller out-of-sample MSE than Model 1 provided that Assumption 1 holds. In other words, provided that the predictor of Model 2 is stronger than local to zero ($\alpha_{x,2} < 1/2$) and more powerful than the predictor of Model 1 ($\alpha_{x,2} < \alpha_{x,1}$), we can establish bounds on the amount by which Model 2 is expected to dominate Model 1. Moreover, the expected MSE gains depend on the strength of the predictive signals, as a stronger predictor (a smaller $\alpha_{x,2}$) is associated with a larger expected gain from using the forecasts from model 2 rather than forecasts from model 1.[17]

3.3 Expected Gains from Forecast Monitoring

In situations where two or more alternative forecasts of the same outcome are available, we can define the gain from monitoring forecasting performance of a particular model as the expected reduction in loss from selecting the other model when this is expected to perform better. We next describe the setup for forecast monitoring and analyze the expected gains from forecast monitoring. We assume that the parameters of the forecast monitoring regression in (6) are estimated using a monitoring window based on the most recent m observations

$$\hat{\theta}_{m,t} = \left(m^{-1} \sum_{s=t-m}^{t-1} z_s z_s' \right)^{-1} \left(m^{-1} \sum_{s=t-m}^{t-1} z_s \Delta L_{s+1} \right), \tag{12}$$

where $z_s = (1, z_{1s})'$.

The switching rule chooses model 2 if and only if $z_t' \hat{\theta}_{m,t} > 0$. Using (8), the forecasts from the monitoring rule $\{\hat{y}_{SW,t+1}\}_{t=n+m}^{T-1}$ take the form

$$\hat{y}_{SW,t+1|t} = \hat{y}_{1,t+1|t} \mathbf{1}\{z_t' \hat{\theta}_{m,t} \leq 0\} + \hat{y}_{2,t+1|t} \mathbf{1}\{z_t' \hat{\theta}_{m,t} > 0\}. \tag{13}$$

[17] Using a Diebold-Mariano test based on a sample of size T, we can detect differences of an order larger than $O(T^{-1/2})$. Hence, by Proposition 1, a Diebold-Mariano test will have power in large samples to detect differences in the average performance of models 1 and 2 if $\alpha_{x,2} < 1/2$. Conversely, for larger values of $\alpha_{x,2}$, and hence for weaker predictors, such tests will not be very powerful.

Under the recursive monitoring rule, the sample size, T, is split into rolling windows of length n and m used to estimate $\hat{\beta}_{j,n,t}$ (the parameters of the jth forecasting model) and $\hat{\theta}_{m,t}$ (the parameters of the monitoring regression), respectively, and an evaluation sample containing the remaining p observations so that $T = n + m + p$. The tests that evaluate the performance of the two models and the switching rule are thus based on the p forecasts $\{(\hat{y}_{1,t+1|t}, \hat{y}_{2,t+1|t}, \hat{y}_{SW,t+1|t})\}_{t=n+m}^{n+m+p-1}$. Again, n, m and p are viewed as functions of T which may or may not tend to infinity and, for simplicity, we write n, m and p, instead of n_T, m_T and p_T. Figure 2 illustrates the setup of our analysis, showing how the observations are split into estimation, monitoring and forecast evaluation samples.

To establish results on the expected gains from forecast monitoring, using (11) we need to make assumptions about the correlation between the monitoring instrument, z_{1t}, and the $x_{i,t}\varepsilon_{t+1}$ terms in the forecast errors. We collect these in Assumption 2:

Assumption 2. *The following hold for $j \in \{1, 2\}$:*

(i) $\{x_{j,t}, z_{1t}, \varepsilon_{t+1}\}_{t=-\infty}^{\infty}$ *is a β-mixing array with coefficient $\beta_{mix}(\cdot)$ such that $\forall t > 0$, $\beta(t) \leq b \exp(-t^c)$, for constants $b, c > 0$.*

(ii) There exist constants $\alpha_{z,1}, \alpha_{z,2} \in [0, \infty]$, $c_{\rho,1}, c_{\rho,2} > 0$ such that $corr(x_{j,t}\varepsilon_{t+1}, z_{1t}) = c_{\rho,j} m^{-\alpha_{z,j}}$, where

(a) $2r\alpha_{z,2}/(r-2) < \alpha_{x,2}$,

(b) $\alpha_{x,2} + \alpha_{z,2} < \alpha_{x,1} + \alpha_{z,1}$.

(iii) For some constants $\kappa_1, \kappa_2 > 0$, $\kappa_1 \mathbb{E}x_{2,t}\varepsilon_{t+1}z_{1t} \leq \mathbb{E}x_{2,t}\varepsilon_{t+1}\mathbf{1}\{z_{1t} > 0\} \leq \kappa_2 \mathbb{E}x_{2,t}\varepsilon_{t+1}z_{1t}$.

(iv) On some fixed neighborhood of zero, the p.d.f. of z_{1t} is uniformly bounded.

(v) $\mathbb{E}z_{1t} = 0$. *Moreover, for constants $r, D > 0$, and $\mathbb{E}|z_{1t}|^r \leq D$.*

(vi) T/m is bounded.

The mixing condition in Assumption 2(i) ensures weak dependence in the predictors, monitoring instrument, and outcomes and so this assumption naturally extends Assumption 1(i). Assumption 2(ii) ensures that the monitoring instrument, z_{1t}, is not

too weak for the second model (part a) and that the "combined" strength of the predictor and monitoring instrument ($\alpha_{x,2}+\alpha_{z,2}$) is stronger for model 2 than for model 1 (part b). Assumption 2(iii) links the selection rule and the correlation between $x_{2,t}\varepsilon_{t+1}$ and z_{1t}. The condition says that the correlation between $x_{2,t}\varepsilon_{t+1}$ and z_{1t} is of the same order of magnitude as the correlation between $x_{2,t}\varepsilon_{t+1}$ and $\mathbf{1}\{z_{1t}>0\}$. This means that the dependence between $x_{2,t}\varepsilon_{t+1}$ and z_{1t} can be measured in approximately equivalent ways, either by the correlation between $x_{2,t}\varepsilon_{t+1}$ and z_{1t} or by the correlation between $x_{2,t}\varepsilon_{t+1}$ and $\mathbf{1}\{z_{1t}>0\}$.[18] Assumptions 2(iv)-(v) impose mild assumptions on the distribution and moments of z_{1t}. We also assume in part (vi) that the length of the monitoring window, m, grows in proportion with the sample size, T.

With Assumptions 1 and 2 in place, we can characterize the expected gains from monitoring forecasting performance, i.e., the expected performance of the switching rule relative to models 1 and 2:

Proposition 2. *Consider the DGP*

$$y_{t+1}=\beta_1 x_{1,t}+\beta_2 x_{2,t}+\varepsilon_{t+1}.$$

Moreover, assume that the parameters of the forecasting models, $\hat{\beta}_{j,n,t}$, $(j=1,2)$ and of the monitoring rule, $\hat{\theta}_{m,t}$, are estimated recursively using n and m observations, respectively. Then, assuming squared error loss, under Assumptions 1 and 2, the following hold:

(1) There exist constants $G_1,G_2,G_3>0$ and an array $\{S_{t+1}\}_{t=n+m}^{T-1}$ such that for $T\geq G_1$ and $n+m\leq t\leq T-1$

$$\mathbb{P}\left(\bigcap_{t=n+m}^{T-1}\left\{S_{t+1}=\Delta L_{t+1}\mathbf{1}\{z_t'\hat{\theta}_{m,t}>0\}\right\}\right)\geq 1-G_2T^{\max\{1-r/8,\ 1+(\alpha_{x,2}-1/2)(2+r/4)\}}$$

[18] Notice that the switching rule aims to predict the sign of ΔL_{t+1} using z_{1t}. However, the success of the switching rule in terms of $\mathbb{E}\Delta L_{t+1}\mathbf{1}\{z_{1t}>0\}$ is not guaranteed merely by a strong correlation between $x_{2,t}\varepsilon_{t+1}$ and z_{1t} even if $x_{2,t}\varepsilon_{t+1}$ is the dominant term of ΔL_{t+1}. To see this, consider the following situation. Let u_t be a random variable with a symmetric distribution around zero. For any $\alpha_x\in(0,1)$, let ξ_t be a Bernoulli random variable that equals $1/(\alpha_x-1)$ with probability $(1-\alpha_x)/2$ and equals $1/(1+\alpha_x)$ with probability $(1+\alpha_x)/2$. Consider $z_{1t}=u_t\xi_t$ and $x_{2,t}\varepsilon_{t+1}=\alpha_x z_{1t}+u_t$. It is not hard to show that $corr(x_{2,t}\varepsilon_{t+1},z_{1t})=\alpha_x$; however, it can be easily shown that $\mathbb{E}x_{2,t}\varepsilon_{t+1}\mathbf{1}\{z_{1t}>0\}=0$ for any $\alpha_x\in(0,1)$.

and

$$\mathbb{E}S_{t+1} \geq G_3 T^{-(\alpha_{x,2}+\alpha_{z,2})}.$$

(2) There exist constants $G_4, G_5, G_6 > 0$ *and an array* $\{\tilde{S}_{t+1}\}_{t=n+m}^{T-1}$ *such that for* $T \geq G_4$ *and* $n+m \leq t \leq T-1$

$$\mathbb{P}\left(\bigcap_{t=n+m}^{T-1}\left\{\tilde{S}_{t+1} = -\Delta L_{t+1}\mathbf{1}\{z_t'\hat{\theta}_{m,t} < 0\}\right\}\right) \geq 1 - G_5 T^{\max\{1-r/8,\ 1+(\alpha_{x,2}-1/2)(2+r/4)\}}$$

and

$$\mathbb{E}\tilde{S}_{t+1} \geq G_6 T^{-(\alpha_{x,2}+\alpha_{z,2})}.$$

To interpret part 1 of Proposition 2, notice that $(y_{t+1} - \hat{y}_{1,t+1|t})^2 - (y_{t+1} - \hat{y}_{SW,t+1|t})^2 = \Delta L_{t+1}\mathbf{1}\{z_t'\hat{\theta}_{m,t} > 0\}$ so S_{t+1} captures the expected gains from monitoring the performance of model 1, i.e., the squared error loss of model 1 relative to the switching rule, provided that $T^{\max\{1-r/8,\ 1+(\alpha_{x,2}-1/2)(2+r/4)\}}$ is small. When this holds, the result shows that tests computed based on $\{\Delta L_{t+1}\mathbf{1}\{z_t'\hat{\theta}_{m,t} > 0\}\}$ have the same asymptotic properties as those computed using S_{t+1}.

By part 1 of Proposition 2, the expected gain from monitoring the performance of the first model is bounded below by a positive sequence of order $T^{-(\alpha_{x,2}+\alpha_{z,2})}$. Hence, the more accurate the predictor variable of model 2 (i.e., the smaller is $\alpha_{x,2}$) and the better the monitoring instrument (smaller $\alpha_{z,2}$), the bigger the expected gain from monitoring the performance of model 1 and switching to use the forecasts from model 2 when this model is expected to have the best performance.

Part 2 of Proposition 2 computes the expected gain from monitoring the performance of Model 2. To see this, notice that $(y_{t+1} - \hat{y}_{2,t+1|t})^2 - (y_{t+1} - \hat{y}_{SW,t+1|t})^2 = -\Delta L_{t+1}\mathbf{1}\{z_t'\hat{\theta}_{m,t} < 0\}$, so $\tilde{S}_{t+1}$ captures the squared error loss of model 2 relative to the switching rule with a high probability if $G_5 T^{\max\{1-r/8,\ 1+(\alpha_{x,2}-1/2)(2+r/4)\}}$ is small. Hence, part 2 of Proposition 2 says that, to a good approximation, the expected loss of the switching rule measured relative to the second model is also bounded below by a positive sequence of order $T^{-(\alpha_{x,2}+\alpha_{z,2})}$.

Proposition 1 established that model 2 is expected to outperform model 1. To see why the expected gain from monitoring the performances of models 1 or 2 are of the same order of magnitude, note that the leading term in the MSE differential of the switching rule versus either model 1 or model 2 in equation (11) is of order

$T^{-(\alpha_{x,2}+\alpha_{z,2})}$ whereas the expected MSE differential of model 2 versus model 1 is of order $T^{-2\alpha_{x,2}}$. Our assumption that $\alpha_{z,2} < \alpha_{x,2}$ therefore ensures that the expected MSE differential of the two models is of a smaller order than $T^{-(\alpha_{x,2}+\alpha_{z,2})}$.

An important condition for the switching rule to work is that $\mathbb{E}[x_{jt}\varepsilon_{t+1}z_{1t}] \neq 0$, so that the monitoring instrument, z_{1t}, is capable of picking up predictable forecast errors. This condition can hold even if $\mathbb{E}[\varepsilon_{t+1}z_{1t}] = 0$. Hence, the instrument need not have any predictive power if added to the forecasting model on its own. Monitoring instruments can therefore be useful for tracking the (relative) expected loss of a particular forecast even though they need not have predictive power over the outcome as stand-alone predictors. Of course, adding the cross-product term $x_{jt}z_{1t}$ as a predictor to the original forecasting model might produce better results. However, this strategy is often not a feasible option since x_t might not be observed, as in the case of survey data or any third-party forecasts that are not generated by the forecast user.[19]

3.4 Weak Predictor with a Strong Monitoring Instrument

Proposition 2 establishes results for the switching rule under conditions ensuring that the predictor and monitoring instrument are more powerful for Model 2 than for Model 1 through the assumptions $\alpha_{x,2} < \alpha_{x,1}$ and $\alpha_{x,2} + \alpha_{z,2} < \alpha_{x,1} + \alpha_{z,1}$. In this case, the dominant term in the monitoring rule is the correlation between z_{1t} and $2\beta_2\varepsilon_{t+1}x_{2,t}$. However, suppose that model 2 uses the strongest predictor ($\alpha_{x,2} < \alpha_{x,1}$) but that the monitoring instrument is stronger for model 1 and that $\alpha_{x,2} + \alpha_{z,2} > \alpha_{x,1} + \alpha_{z,1}$. In this case, the dominant term in the monitoring rule becomes the correlation between z_{1t} and $2\beta_1\varepsilon_{t+1}x_{1,t}$. We next show that it is possible to generate gains from forecast monitoring also in this case. We capture the case with a weak predictor and a strong monitoring instrument through the following assumption:

Assumption 3. *Let Assumption 2 (i), (iii)-(vi) hold for some $r \geq 10$, but replace Assumption 2(ii) with the assumption that there exist constants $\alpha_{z,1}, \alpha_{z,2} \in [0, \infty]$, $c_{\rho,1}, c_{\rho,2} > 0$ such that $corr(x_{j,t}\varepsilon_{t+1}, z_{1t}) = c_{\rho,j}m^{-\alpha_{z,j}}$ and*

[19] Note that if $r > \max\{8,\ 16\alpha_{x,2}/(1-2\alpha_{x,2})\}$, then $T^{\max\{1-r/8,\ 1+(\alpha_{x,2}-1/2)(2+r/4)\}}$ vanishes and tests computed based on $\{\Delta L_{t+1}\mathbf{1}\{z_t'\hat{\theta}_{m,t} > 0\}\}$ have the same asymptotic properties as those computed using $\{S_{t+1}\}$. Hence, using a Diebold-Mariano test and a sample of size T, we can detect differences in forecasting performance of an order larger than $O(T^{-1/2})$. By Proposition 2, differences between the switching rule and models 1 or 2 can therefore be detected if $\alpha_{x,2} + \alpha_{z,2} < 1/2$.

1. $\alpha_{z,1} < \alpha_{x,1}$,

2. $\alpha_{x,1} + \alpha_{z,1} < \min\{1/2,\ (3r-2)\alpha_{x,2}/(2r-2)\}$, *and*

3. $\alpha_{x,2} + \alpha_{z,2} > \alpha_{x,1} + \alpha_{z,1}$.

Note the new parameter restrictions for this case. We require that the monitoring instrument be more strongly correlated with the cross-product $x_{1,t}\varepsilon_{t+1}$ than the correlation between the "weak" predictor and the outcome ($\alpha_{z,1} < \alpha_{x,1}$), at least for large T. We also require that the sum $\alpha_{x,1} + \alpha_{z,1} < 1/2$, although this bound could be tighter, depending on the values of r and $\alpha_{x,2}$. The last part of Assumption 3 captures that the combined strength of the predictor and monitoring instrument for model 1 is stronger than that for model 2.

Proposition 3. *Suppose that Assumptions 1 and 3 are satisfied. Then, the following hold:*

(1) There exist constants $G_1, G_2, G_3 > 0$ and an array $\{S_{t+1}\}_{t=n+m}^{T-1}$ such that for $T \geq G_1$ and $n+m \leq t \leq T-1$

$$\mathbb{P}\left(\bigcap_{t=n+m}^{T-1}\left\{S_{t+1} = \Delta L_{t+1}\mathbf{1}\{z_t'\hat{\theta}_{m,t} > 0\}\right\}\right) \geq 1 - G_2 T^{\max\{1-r/8,\ 1+(\alpha_{x,2}-1/2)(2+r/4)\}}$$

and

$$\mathbb{E}S_{t+1} \geq G_3 T^{-\alpha_{x,1}-\alpha_{z,1}}.$$

(2) There exist constants $G_4, G_5, G_6 > 0$ and an array $\{\tilde{S}_{t+1}\}_{t=n+m}^{T-1}$ such that for $T \geq G_4$ and $n+m \leq t \leq T-1$

$$\mathbb{P}\left(\bigcap_{t=n+m}^{T-1}\left\{\tilde{S}_{t+1} = -\Delta L_{t+1}\mathbf{1}\{z_t'\hat{\theta}_{m,t} < 0\}\right\}\right) \geq 1 - G_5 T^{\max\{1-r/8,\ 1+(\alpha_{x,2}-1/2)(2+r/4)\}}$$

and

$$\mathbb{E}\tilde{S}_{t+1} \geq G_6 T^{-(\alpha_{x,1}+\alpha_{z,1})}.$$

Proposition 3 shows that the expected gain from monitoring forecasting performance relative to either always using the forecasts from model 1 or always using the forecasts from model 2 is bounded from below by terms that are of order $T^{-(\alpha_{x,1}+\alpha_{z,1})}$,

which, by assumption, is bigger than $T^{-1/2}$. Together with the result in Proposition 2, this shows that we can have expected gains from monitoring either in situations where we have a strong predictor and a monitoring instrument that is strongly correlated with the cross-product of this predictor and the residual from the forecasting model (Proposition 2) or in cases with a weak predictor but a monitoring instrument that is strongly correlated with the cross-product of the weak predictor and the residual from the forecasting model (Proposition 3).

4 Nested Models

Comparisons of forecasts from nested models arise in a number of applications in economics and finance and this case can be addressed by modifying the analysis in Section 3. Suppose the data generating process (DGP) includes an intercept and a time varying regressor and thus takes the form

$$y_{t+1} = \mu + \beta x_t + \varepsilon_{t+1}, \tag{14}$$

Moreover, suppose that model 2 (the "big" model) coincides with the DGP in (14), while model 1 is a (nested) small model that only includes an intercept:

$$y_{t+1} = \mu + \varepsilon_{t+1}. \tag{15}$$

This setup captures a number of economically interesting cases, including attemps to capture time-varying predictability of asset returns measured relative to a constant expected returns benchmark.

We estimate both models using OLS so that, for $n \leq t \leq T$,

$$\begin{aligned} \tilde{\mu}_t &= n^{-1} \sum_{s=t-n+1}^{t} y_s, \\ \begin{pmatrix} \hat{\mu}_t \\ \hat{\beta}_t \end{pmatrix} &= \left[n^{-1} \sum_{s=t-n}^{t-1} \begin{pmatrix} 1 \\ x_s \end{pmatrix} \begin{pmatrix} 1 & x_s \end{pmatrix} \right]^{-1} \left[n^{-1} \sum_{s=t-n}^{t-1} \begin{pmatrix} 1 \\ x_s \end{pmatrix} y_{s+1} \right], \end{aligned} \tag{16}$$

and $\hat{y}_{1,t+1|t} = \tilde{\mu}_t$, while $\hat{y}_{2,t+1|t} = \hat{\mu}_t + \hat{\beta}_t x_t$.

Using these notations, the difference in squared error losses is

$$\begin{aligned}\Delta L_{t+1} &= (y_{t+1} - \tilde{\mu}_t)^2 - (y_{t+1} - \hat{\mu}_t - \hat{\beta}_t x_t)^2 \\ &= (-\delta_{t,small} + \beta x_t + \varepsilon_{t+1})^2 - (-\delta_{t,big} + \varepsilon_{t+1})^2 \\ &= \beta^2 x_t^2 + 2\beta x_t \varepsilon_{t+1} + \delta_{t,small}^2 - \delta_{t,big}^2 + 2\delta_{t,big}\varepsilon_{t+1} - 2\delta_{t,small}(\beta x_t + \varepsilon_{t+1}), \quad (17)\end{aligned}$$

where $\delta_{t,small} = \tilde{\mu}_t - \mu$ and $\delta_{t,big} = \hat{\mu}_t - \mu + (\hat{\beta}_t - \beta)x_t$.

A particularly simple case with nested models arises when the DGP and forecasting models do not include an intercept. For this case (17) simplifies to

$$\Delta L_{t+1} \equiv (y_{t+1} - 0)^2 - \left(y_{t+1} - \hat{\beta}_{n,t} x_t\right)^2 = (\beta^2 - \delta_{\beta,n,t}^2)x_t^2 + 2(\beta + \delta_{\beta,n,t})x_t\varepsilon_{t+1}. \quad (18)$$

Using the earlier notations, we can capture this case by setting $\alpha_{x,1} = \alpha_{z,1} = \infty$, so that $\beta_1 = \delta_{1,n,t} = 0$. This allows us to simplify the notations by setting $\beta_2 = \beta = cn^{-\alpha_x}$ and $corr(x_t\varepsilon_{t+1}, z_{1t}) = c_\rho n^{-\alpha_z}$, where $c, c_\rho > 0$ and $\alpha_x, \alpha_z \geq 0$ are constants. Moreover, $\delta_{\beta,2,n,t} = \delta_{\beta,n,t}$, where $\delta_{\beta,n,t} = (\sum_{s=t-n}^{t-1} x_s^2)^{-1}(\sum_{s=t-n}^{t-1} x_s\varepsilon_{s+1})$.

Note a subtle difference between the nested and non-nested case: In the nested case, we impose on the small model that the parameter of the additional predictor that is only included in the big model takes a value of zero so that fewer parameters are estimated by the small model. Conversely, in the non-nested case, no such constraint is imposed and so we do not have a "big" and a "small" model for this case.

4.1 Expected Performance of Big versus Small Forecasting Models

We summarize our list of assumptions for the case with nested models in Assumption 4:

Assumption 4. *Assume that the following hold*
(i) The r-th moments of x_t, z_{1t} and ε_{t+1} are uniformly bounded for some constant $r > 8$.
(ii) $\{x_t, z_{1t}, \varepsilon_t\}_{t=-\infty}^{\infty}$ is a β-mixing array with coefficient $\beta_{mix}(\cdot)$ such that $\forall t > 0$, $\beta_{mix}(t) \leq b\exp(-t^c)$, for constants $b, c > 0$.
(iii) $\mathbb{E}(\varepsilon_{t+1} \mid \{(x_s, \varepsilon_s)\}_{s=-\infty}^{t}) = 0$ and $\mathbb{E}x_t = \mathbb{E}z_{1t} = 0$.
(iv) $\beta = cn^{-\alpha_x}$ for some constants $\alpha_x \in [0, \infty)$, $c > 0$.

(iv) $M_1 \leq \mathbb{E}x_t^2 \leq M_2$ *for some constants* $M_1, M_2 > 0$.
(v) T/n *and* T/m *are bounded.*

Using this assumption, we can characterize the expected squared error loss performance of the small versus the big models for the nested case:

Proposition 4. *Consider the data generating process*

$$y_{t+1} = \mu + x_t\beta + \varepsilon_{t+1},$$

and suppose that Assumption 4 holds.

(1) Suppose that $\alpha_x < 1/2$. *Then there exist constants* $C_1, \cdots, C_4 > 0$ *and an array* $\{\Delta L_{t+1,*}\}_{t=n}^{T-1}$ *such that for* $T \geq C_1$

$$\mathbb{P}\left(\bigcap_{t=n+m}^{T-1} \{\Delta L_{t+1} = \Delta L_{t+1,*}\}\right) \geq 1 - C_2 T^{\max\{1-r/8,\ 1+(\alpha_x-1/2)(2+r/4)\}}$$

and

$$C_3 T^{-2\alpha_x} \leq \mathbb{E}\Delta L_{t+1,*} \leq C_4 T^{-2\alpha_x} \quad \forall n \leq t \leq T-1.$$

(2) Suppose that $\alpha_x > 1/2$. *Then there exist constants* $C_5, \cdots, C_8 > 0$ *and an array of random variables* $\{\Delta L_{t+1,*}\}_{t=n}^{T-1}$ *such that for* $T \geq C_5$

$$\mathbb{P}\left(\bigcap_{t=n+m}^{T-1} \{\Delta L_{t+1} = \Delta L_{t+1,*}\}\right) \geq 1 - C_6 T^{\max\{1-r/8,\ 1+(\alpha_x-1/2)(2+r/4)\}}$$

and

$$-C_7 T^{-1} \leq \mathbb{E}\Delta L_{t+1,*} \leq -C_8 T^{-1} \quad \forall n \leq t \leq T-1.$$

Part 1 of Proposition 4 shows that the expected squared error loss of the big model is smaller than that of the small model that excludes this predictor provided that the strength of the predictor included in the big model is sufficiently large to overcome the effect of estimation error ($\alpha_x < 1/2$). Moreover, the amount by which the big model is expected to outperform the small model gets bigger, the stronger the predictive signal, i.e., the smaller is α_x. Conversely, part 2 of Proposition 4 says that if the predictive signal underlying the big model is "too weak" ($\alpha_x > 1/2$), then the estimation error of the big model dominates the signal, leading us to expect that the big model will

underperform the small model, although the expected underperformance is only of order $O(T^{-1})$.

Next, consider whether a statistical test will have the power to differentiate between the predictive accuracy of the big and the small model. From equation (17), it is not hard to see that $\mathbb{E}(\Delta L_{t,n,*})^2$ is roughly of order $O(n^{-2\alpha_x})$. Assuming that p and n are of the same order, this means that for $\alpha_x < 1/2$, $T^{1/2}\mathbb{E}\Delta L_{t,n,*}/\sqrt{\mathbb{E}(\Delta L_{t,n,*})^2}$ is roughly of order $O(T^{1/2-\alpha_x})$, which tends to infinity. Hence, a Diebold-Mariano t-test or a GW test will tend to reject $\mathbb{E}\left[\Delta L_{t,n,*}\right] = 0$ if $\alpha_x < 1/2$. In the case with a weak predictor in the large model ($\alpha_x > 1/2$), the usual Diebold-Mariano test using p observations of ΔL_{t+1} will not detect such underperformance when $T/p = O(1)$ because the expected underperformance is only of order $O(T^{-1})$.

4.2 Expected Gains from Monitoring Forecasting Performance

As in the case with non-nested models, we next characterize the behavior of the switching rule for the nested case. For this analysis we make use of the following assumption:

Assumption 5. *The following hold*

(i) There exist constants $\alpha_z \in [0, \infty]$, $c_\rho > 0$ such that $corr(x_t\varepsilon_{t+1}, z_{1t}) = c_\rho m^{-\alpha_z}$, where $2r\alpha_z/(r-2) < \alpha_x$.

(ii) For some constants $\kappa_1, \kappa_2 > 0$, $\kappa_1 \mathbb{E}x_t\varepsilon_{t+1}z_{1t} \leq \mathbb{E}x_t\varepsilon_{t+1}\mathbf{1}\{z_{1t} > 0\} \leq \kappa_2\mathbb{E}x_t\varepsilon_{t+1}z_{1t}$.

(iii) On some fixed neighborhood of zero, the p.d.f. of z_{1t} is uniformly bounded.

Using Assumption 5, we have the following result for the switching rule in the nested case:

Proposition 5. *Consider the data generating process*

$$y_{t+1} = \mu + x_t\beta + \varepsilon_{t+1}.$$

Suppose Assumptions 4 and 5 hold. Then

(1) there exist constants $M_1, M_2, M_3 > 0$ *and an array* $\{S_{t+1}\}_{t=n+m}^{T-1}$ *such that for* $T \geq M_1$ *and* $\forall n+m \leq t \leq T-1$,

$$\mathbb{P}\left(\bigcap_{t=n+m}^{n+m+p-1}\left\{S_{t+1} = \Delta L_{t+1}\mathbf{1}\{z_t'\hat{\theta}_{m,t} > 0\}\right\}\right) \geq 1 - M_2 T^{\max\{1-r/8,\ 1+(\alpha_x-1/2)(2+r/4)\}}$$

and

$$\mathbb{E}S_{t+1} \geq M_3 T^{-(\alpha_x+\alpha_z)}.$$

(2) there exist constants $M_4, M_5, M_6 > 0$ *and an array*$\{\tilde{S}_{t+1}\}_{t=n+m}^{T-1}$ *such that for* $T \geq M_4$ *and* $\forall n+m \leq t \leq T-1$,

$$\mathbb{P}\left(\bigcap_{t=n+m}^{n+m+p-1}\left\{\tilde{S}_{t+1} = -\Delta L_{t+1}\mathbf{1}\{z_t'\hat{\theta}_{m,t} < 0\}\right\}\right) \geq 1 - M_5 T^{\max\{1-r/8,\ 1+(\alpha_x-1/2)(2+r/4)\}}$$

and

$$\mathbb{E}\tilde{S}_{t+1} \geq M_6 T^{-(\alpha_x+\alpha_z)}.$$

Part 1 of Proposition 5 shows that the switching rule is expected to perform better than the small model by an amount that is bounded by a factor of order $T^{-(\alpha_x+\alpha_z)}$. A similar result holds for the amount by which the switching rule is expected to outperform the big model.

4.2.1 Effect of Persistent Estimation Errors

Under the rolling estimation scheme, estimation errors are highly persistent and, thus, predictable by means of their lagged values. Under some conditions, it is possible to utilize this fact and build a switching rule that generates superior performance.

As an illustration, consider the simple DGP $y_{t+1} = \varepsilon_{t+1}$ with a serially correlated error, $\varepsilon_{t+1} = \rho_\varepsilon \varepsilon_t + u_{\varepsilon,t+1}$. Suppose that the big model takes the form $y_{t+1} = \beta x_t + \varepsilon_{t+1}$, where $x_{t+1} = \rho_x x_t + u_{x,t+1}$. Further, assume that $u_{\varepsilon,t} \sim i.i.d\ N(0, \sigma_{u,\varepsilon}^2)$ is independent of $u_{x,t} \sim i.i.d\ N(0, \sigma_{u,x}^2)$. Because the true population value of β is zero, the loss differential is given by

$$\Delta L_{t+1} \quad = \quad \delta_{\beta,n,t}^2 x_t^2 + 2\delta_{\beta,n,t} x_t \varepsilon_{t+1}. \tag{19}$$

In cases with a fixed t and a large n, $\delta_{\beta,n,t} = O_P(n^{-1/2})$ and so $\Delta L_{t+1} =$

$O_P(n^{-1}) + 2\delta_{\beta,n,t}x_t\varepsilon_{t+1}$. Thus, $\mathbb{E}\Delta L_{t+1} = O(n^{-1})$, $\mathbb{E}(\Delta L_{t+1})^2 = O(n^{-2}) + 4\mathbb{E}(\delta^2_{\beta,n,t}x_t^2)\mathbb{E}(\varepsilon^2_{t+1})$, and $\mathbb{E}(\Delta L_{t+1}\Delta L_{t+2}) = O(n^{-1}) + 4\mathbb{E}(\delta_{\beta,n,t}\delta_{\beta,n,t+1}x_t x_{t+1}\varepsilon_{t+1}\varepsilon_{t+2})$. Assuming weak dependence of the data, $(\delta_{\beta,n,t}, \delta_{\beta,n,t+1})$ is asymptotically independent of $(x_t, x_{t+1}, \varepsilon_{t+1}, \varepsilon_{t+2})$.[20] Hence, we can compute (a) $\mathbb{E}(\Delta L_{t+1})^2 = O(n^{-2}) + 4\mathbb{E}(\delta^2_{\beta,n,t})\mathbb{E}(x_t^2)\mathbb{E}(\varepsilon^2_{t+1})$ and (b) $cov(\Delta L_{t+1,n}, \Delta_{t+2,n}) = O(n^{-1}) + 4\mathbb{E}(\delta_{\beta,n,t}\delta_{\beta,n,t+1})\mathbb{E}(x_t x_{t+1}\varepsilon_{t+1}\varepsilon_{t+2}) = O(n^{-1}) + 4\mathbb{E}(\delta_{\beta,n,t}\delta_{\beta,n,t+1})\rho_x\mathbb{E}(x_t^2)\rho_\varepsilon\mathbb{E}(\varepsilon^2_{t+1})$. Now (a) and (b), together with $\mathbb{E}\Delta L_{t+1} = O(n^{-1})$, imply that $corr(\Delta L_{t+1}, \Delta L_{t+2}) = \rho_x\rho_\varepsilon corr(\delta_{\beta,n,t}, \delta_{\beta,n,t+1}) + o(1)$. Since $corr(\delta_{\beta,n,t}, \delta_{\beta,n,t+1}) \to 1$, we have

$$corr(\Delta L_{t+1}, \Delta L_{t+2}) \to \rho_x\rho_\varepsilon.$$

This observation yields a simple monitoring rule: if $\rho_x\rho_\varepsilon > 0$, then choose the big model at time t when $\Delta L_t > 0$; if $\rho_x\rho_\varepsilon < 0$, then choose the big model at time t when $\Delta L_t < 0$.[21]

5 Results in the Absence of Estimation Error

The analysis in the previous two sections is complicated by the presence of estimation error in the models used to generate forecasts. To gain intuition and simplify the analysis, this section considers the case without estimation error. In addition, we consider the case with a highly persistent (near unit root) predictor and go on to discuss the choice of monitoring instruments and possible sources of gains from monitoring forecasting performance which become particularly transparent in the absence of estimation error.

[20] To see this, we provide an intuitive argument in the spirit of Bernstein's block technique. Consider $\delta_{\beta,n,t-k_n}$ for some $k_n \to \infty$ but $k_n/n \to 0$. Since the data used to compute $\delta_{\beta,n,t}$, $\delta_{\beta,n,t+1}$ and $\delta_{\beta,n,t-k_n}$ mostly overlap ($k_n/n = o(1)$); they all contain $\{(x_s, y_{s+1}) \mid t-n+2 \le s \le t-k_n-1\}$. Thus, $\delta_{\beta,n,t} = (1+o_P(1))\delta_{\beta,n,t-k_n}$ and $\delta_{\beta,n,t+1} = (1+o_P(1))\delta_{\beta,n,t-k_n}$. On the other hand, due to the weak dependence, $\delta_{\beta,n,t-k_n}$ and $(x_t, x_{t+1}, \varepsilon_{t+1}, \varepsilon_{t+2})$ are asymptotically independent as $k_n \to \infty$. See Bradley (2007) for formal arguments.

[21] To determine the size of the correlation, one can simply observe that $2\delta_{\beta,n,t}x_t\varepsilon_{t+1}$ is the leading term in (19) and estimate the autocorrelation of $x_t\hat{\varepsilon}_{t+1}$, where $\hat{\varepsilon}_{t+1}$ is the regression residual from the big model.

5.1 Non-nested case

In the absence of estimation error, equation (11) simplifies to

$$\begin{aligned}\Delta L_{t+1} &= (\beta_2 x_{2,t} + \varepsilon_{t+1})^2 - (\beta_1 x_{1,t} + \varepsilon_{t+1})^2 \\ &= \left(\beta_2^2 x_{2,t}^2 - \beta_1^2 x_{1,t}^2\right) + 2\left(\beta_2 x_{2,t} - \beta_1 x_{1,t}\right)\varepsilon_{t+1}. \end{aligned} \tag{20}$$

Thus, to forecast ΔL_{t+1}, an instrument must be able to predict $x_{1,t}^2$, $x_{2,t}^2$, $x_{1,t}\varepsilon_{t+1}$ or $x_{2,t}\varepsilon_{t+1}$.

Defining $W_t = \mathbb{E}[\Delta L_{t+1}|z_t]$, it suffices to bound $\mathbb{E}\left[W_t \mathbf{1}\{W_t > 0\}\right]$ and $\mathbb{E}\left[W_t \mathbf{1}\{W_t \leq 0\}\right]$. In the absence of estimation error, $\mathbb{E}\left[W_t\right] = \beta_2^2 \mathbb{E}x_{2,t}^2 - \beta_1^2 \mathbb{E}x_{1,t}^2$ and

$$|W_t| \geq 2\left|\mathbb{E}\left[\left(\beta_2 x_{2,t} - \beta_1 x_{1,t}\right)\varepsilon_{t+1} \mid z_t\right]\right| - \left|\mathbb{E}[(\beta_2^2 x_{2,t}^2 - \beta_1^2 x_{1,t}^2) \mid z_t]\right|. \tag{21}$$

Using that $W_t \mathbf{1}\{W_t > 0\} = (W_t + |W_t|)/2$, we have[22]

$$\begin{aligned}&\mathbb{E}\left[W_t \mathbf{1}\{W_t > 0\}\right] \\ &\geq \frac{2\mathbb{E}\left|\mathbb{E}\left[\left(\beta_2 x_{2,t} - \beta_1 x_{1,t}\right)\varepsilon_{t+1} \mid z_t\right]\right| - \mathbb{E}\left|\mathbb{E}[(\beta_2^2 x_{2,t}^2 - \beta_1^2 x_{1,t}^2) \mid z_t]\right| + \beta_2^2 \mathbb{E}x_{2,t}^2 - \beta_1^2 \mathbb{E}x_{1,t}^2}{2} \\ &\geq \mathbb{E}\left|\mathbb{E}\left[\left(\beta_2 x_{2,t} - \beta_1 x_{1,t}\right)\varepsilon_{t+1} \mid z_t\right]\right| + \min\left\{0, \beta_2^2 \mathbb{E}x_{2,t}^2 - \beta_1^2 \mathbb{E}x_{1,t}^2\right\},\end{aligned}$$

Since $\mathbb{E}\left[W_t \mathbf{1}\{W_t > 0\}\right] \geq 0$, we get the following bound on the amount by which the switching rule is expected to outperform model 1:

$$\mathbb{E}\left[W_t \mathbf{1}\{W_t > 0\}\right] \geq \max\left\{0, \mathbb{E}\left|\mathbb{E}\left[\left(\beta_2 x_{2,t} - \beta_1 x_{1,t}\right)\varepsilon_{t+1} \mid z_t\right]\right| + \min\left\{0, \beta_2^2 \mathbb{E}x_{2,t}^2 - \beta_1^2 \mathbb{E}x_{1,t}^2\right\}\right\}.$$

Similar arguments apply in deriving a bound for the switching rule versus model 2. Notice that $-W_t \mathbf{1}\{W_t \leq 0\} = (|W_t| - W_t)/2$, so (21) implies that

$$\begin{aligned}&-\mathbb{E}\left[W_t \mathbf{1}\{W_t \leq 0\}\right] \\ &\geq \frac{2\mathbb{E}\left|\mathbb{E}\left[\left(\beta_2 x_{2,t} - \beta_1 x_{1,t}\right)\varepsilon_{t+1} \mid z_t\right]\right| - \mathbb{E}\left|\mathbb{E}[(\beta_2^2 x_{2,t}^2 - \beta_1^2 x_{1,t}^2) \mid z_t]\right| - \left(\beta_2^2 \mathbb{E}x_{2,t}^2 - \beta_1^2 \mathbb{E}x_{1,t}^2\right)}{2} \\ &\geq \mathbb{E}\left|\mathbb{E}\left[\left(\beta_2 x_{2,t} - \beta_1 x_{1,t}\right)\varepsilon_{t+1} \mid z_t\right]\right| + \min\left\{0, \beta_1^2 \mathbb{E}x_{1,t}^2 - \beta_2^2 \mathbb{E}x_{2,t}^2\right\}.\end{aligned}$$

[22] The last inequality follows from $\mathbb{E}|X| \geq |\mathbb{E}X| = \max\{\mathbb{E}X, -\mathbb{E}X\}$ with $X = \mathbb{E}[(\beta_2^2 x_{2,t}^2 - \beta_1^2 x_{1,t}^2) \mid z_t]$.

We summarize these results in the following proposition.

Proposition 6. *Assuming that all expectations exist, in the absence of estimation error the following bounds on the squared error loss apply to the case with non-nested models and a data generating process (9):*

(i) $\mathbb{E}\left[\Delta L_{t+1}\mathbf{1}\{\mathbb{E}[\Delta L_{t+1}|z_t] > 0\}\right] \geq \max\Big\{0, \mathbb{E}\left|\mathbb{E}\left[\left(\beta_2 x_{2,t} - \beta_1 x_{1,t}\right)\varepsilon_{t+1} \mid z_t\right]\right| + \min\left\{0, \beta_2^2\mathbb{E}x_{2,t}^2 - \beta_1^2\mathbb{E}x_{1,t}^2\right\}\Big\}.$

(ii) $\mathbb{E}\left[-\Delta L_{t+1}\mathbf{1}\{\mathbb{E}[\Delta L_{t+1}|z_t] \leq 0\}\right] \geq \max\Big\{0, \mathbb{E}\left|\mathbb{E}\left[\left(\beta_2 x_{2,t} - \beta_1 x_{1,t}\right)\varepsilon_{t+1} \mid z_t\right]\right| + \min\left\{0, \beta_1^2\mathbb{E}x_{1,t}^2 - \beta_2^2\mathbb{E}x_{2,t}^2\right\}\Big\}.$

Proposition 6 says that if $\beta_2^2\mathbb{E}x_{2,t}^2 - \beta_1^2\mathbb{E}x_{1,t}^2$ is small and the instrument is informative for $\left(\beta_2 x_{2,t} - \beta_1 x_{1,t}\right)\varepsilon_{t+1}$, then positive lower bounds can be established on the expected squared error performance of the switching rule relative to models 1 and 2.

5.2 Nested case

In the absence of estimation error, for the simple case in (18) the loss differential is given by

$$\Delta L_{t+1} = (y_{t+1} - 0)^2 - (y_{t+1} - \beta x_t)^2 = \beta^2 x_t^2 + 2\beta x_t \varepsilon_{t+1}. \tag{22}$$

Since $\mathbb{E}\left[\Delta L_{t+1}\right] = \mathbb{E}\left[\beta^2 x_t^2\right]$, it is better in expectation to always use the big model than always using the small model. This is intuitive given the assumption that we know the true value of β and any signal (x_t) will make the big model outperform the small model. However, even in this ideal case, it is still possible that the switching rule can beat both the big and the small models. The reason is that the big model beats the small model only on average and fluctuation in the term $2\beta x_t\varepsilon_{t+1}$ makes it possible for the small model to sometimes generate forecast with lower expected loss conditional on information in the monitoring instruments. For example, a large negative value of $\beta x_t\varepsilon_{t+1}$ is associated with underperformance for the big model and if some variable can predict when $\beta x_t\varepsilon_{t+1}$ is likely to be negative, then this variable can be used as a monitoring instrument in the switching rule.

To see by how much the switching rule is expected to outperform the small model,

note that

$$
\begin{aligned}
&\mathbb{E}\left[\Delta L_{t+1}\mathbf{1}\left\{\mathbb{E}(\Delta L_{t+1}\mid z_t)\right\}\right]\\
&=\mathbb{E}\left[W_t\mathbf{1}\{W_t>0\}\right]\\
&=\frac{\mathbb{E}\left[W_t\mathbf{1}\{W_t>0\}+W_t\mathbf{1}\{W_t\le 0\}\right]+\mathbb{E}\left[W_t\mathbf{1}\{W_t>0\}-W_t\mathbf{1}\{W_t\le 0\}\right]}{2}\\
&=\frac{\mathbb{E}W_t+\mathbb{E}|W_t|}{2}\\
&\ge \mathbb{E}\left|\,|\beta|\cdot|\mathbb{E}(x_t\varepsilon_{t+1}\mid z_t)|-\frac{1}{2}\beta^2\mathbb{E}(x_t^2\mid z_t)\right|+\frac{1}{2}\beta^2\mathbb{E}x_t^2\\
&\ge \max\left\{|\beta|\cdot\mathbb{E}\left|\mathbb{E}(x_t\varepsilon_{t+1}\mid z_t)\right|,\beta^2\mathbb{E}x_t^2-|\beta|\cdot\mathbb{E}\left|\mathbb{E}(x_t\varepsilon_{t+1}\mid z_t)\right|\right\},
\end{aligned}
\tag{23}
$$

where the first inequality follows from $\mathbb{E}W_t=\beta^2\mathbb{E}x_t^2$ and $|W_t|\ge 2|\beta|\cdot|\mathbb{E}(x_t\varepsilon_{t+1}\mid z_t)|-\beta^2\mathbb{E}(x_t^2\mid z_t)$.

Since $\mathbb{E}[\Delta L_{t+1}]=\beta^2\mathbb{E}x_t^2$, equation (23) also tells us by how much the switching rule is expected to outperform the big model. In the absence of estimation error, the switching rule cannot underperform the big model, and so

$$
\mathbb{E}\left[-\Delta L_{t+1}\mathbf{1}\left\{\mathbb{E}(\Delta L_{t+1}\mid z_t)\le 0\right\}\right]\ge\max\left\{|\beta|\cdot\mathbb{E}\left|\mathbb{E}(x_t\varepsilon_{t+1}\mid z_t)\right|-\beta^2\mathbb{E}x_t^2,0\right\}. \tag{24}
$$

We summarize these computations in the following result:

Proposition 7. *Assuming that all expectations exist, in the absence of estimation error the following bounds hold under squared error loss for the nested case*
(i) $\mathbb{E}\left[\Delta L_{t+1}\mathbf{1}\left\{\mathbb{E}(\Delta L_{t+1}\mid z_t)>0\right\}\right]\ge\max\left\{|\beta|\cdot\mathbb{E}\left|\mathbb{E}(x_t\varepsilon_{t+1}\mid z_t)\right|,\beta^2\mathbb{E}x_t^2-|\beta|\cdot\mathbb{E}\left|\mathbb{E}(x_t\varepsilon_{t+1}\mid z_t)\right|\right\}$.
(ii) $\mathbb{E}\left[-\Delta L_{t+1}\mathbf{1}\left\{\mathbb{E}(\Delta L_{t+1}\mid z_t)\le 0\right\}\right]\ge\max\left\{|\beta|\cdot\mathbb{E}\left|\mathbb{E}(x_t\varepsilon_{t+1}\mid z_t)\right|-\beta^2\mathbb{E}x_t^2,0\right\}$.

Part (i) of Proposition 7 establishes a lower bound on the amount by which the switching rule is expected to outperform the small forecasting model. The switching rule is expected to perform better than the small model in cases where $|\beta|$ and $\mathbb{E}\left|\mathbb{E}(x_t\varepsilon_{t+1}\mid z_t)\right|$ are large because an option for the switching rule is to always choose the big model. Hence, if the big model is much better than the small model (large $|\beta|$ or large $\beta^2\mathbb{E}x_t^2$), we should expect the switching rule to dominate the small forecasting model by always choosing the big model. If, in addition, $\mathbb{E}\left|\mathbb{E}(x_t\varepsilon_{t+1}\mid z_t)\right|$ is large, then z_t provides an accurate signal on when to use either the small or the large model and so we would expect the switching rule to dominate the small model.

Part (ii) of Proposition 7 shows that large values of both $|\beta|$ and $\mathbb{E}\,|\mathbb{E}(x_t\varepsilon_{t+1} \mid z_t)|$ allow the switching rule to beat the big forecasting model by a bigger margin provided that $|\beta| \cdot \mathbb{E}\,|\mathbb{E}(x_t\varepsilon_{t+1} \mid z_t)|$ does not get dominated by $\beta^2\mathbb{E}x_t^2$. For example, if $|\beta|$ is close to zero, the small and large forecasting models generate very similar forecasts and there is not much scope for switching between the two models. Similarly, if $\mathbb{E}\,|\mathbb{E}(x_t\varepsilon_{t+1} \mid z_t)|$ is small, then z_t does not provide accurate information on when the small model might outperform the big model and so always choosing the big model becomes the preferred choice.

5.3 Persistent Regressors

In many empirical applications, the predictor variables are highly persistent. Cases with highly persistent predictor variables can be captured by using a local-to-unity framework for the predictor, i.e., by modeling x_t as an AR(1) process with an autoregressive parameter that is close to unity. The effect of estimation error for this case is quite complicated, so we abstract from estimation error and focus on the case with nested models.

We collect the assumptions required for the case with a persistent regressor in Assumption 6.

Assumption 6. *The following conditions hold for the nested case with a highly persistent predictor*
(i) The predictor x_t is generated by an AR(1) process

$$x_t = \phi x_{t-1} + u_t,$$

where $\{u_t\}_{t=1}^T$ is an independent sequence with $\mathbb{E}u_t = 0$.
(ii) the array $\{(u_t, \varepsilon_t, z_{1t})\}_{t=1}^n$ is strong mixing with mixing coefficient $\alpha_{mix}(\cdot)$.
(iii) $\phi = \exp(-c_\phi/T)$ for some $c_\phi > 0$.
(iv) $\beta = c_\beta T^{-\alpha_x}$ for $\alpha_x > 0$.
(v) $C_1 \leq \mathbb{E}|u_t|^{2+C_0} \leq C_2$ for some constants $C_1, C_2 > 0$.

In the absence of estimation error, using Assumption 6 we can show the following result:

Proposition 8. *Consider the data generating process*

$$y_{t+1} = x_t\beta + \varepsilon_{t+1}$$

Suppose that Assumption 6 holds. Then

(1) there exist constants $M_1, M_2 > 0$ such that $M_1 T^{1-2\alpha_x} \leq \mathbb{E}\Delta L_{t+1} \leq M_2 T^{1-2\alpha_x}$.

(2) there exist constants $M_3, ..., M_6 > 0$ such that

$$\begin{aligned}
&\mathbb{P}\left(\mathbb{E}(\Delta L_{t+1} \mid z_{1t}) \leq 0\right) \\
&\leq T^{-1} \min_{1\leq G\leq t} \left(M_3 G + M_4\sqrt{G(t-G)} + M_5[\alpha(G)]^{C_0/(C_0+2)}(t-G)\right) \\
&\quad + M_6 T^{\alpha_x - 1} \sum_{i=0}^{t-1} [\alpha_{mix}(i)]^{(C_0+1)/(C_0+2)}.
\end{aligned}$$

(3) there exist constants $M_7, M_8 > 0$ such that

$$\begin{aligned}
&\left|\frac{\mathbb{E}\Delta L_{t+1}\mathbf{1}\{\mathbb{E}(\Delta L_{t+1} \mid z_{1t}) > 0\}}{\mathbb{E}\Delta L_{t+1}} - 1\right| \\
&\leq T^{-1} M_7 \min_{1\leq G\leq t} \left(M_3 G + M_4\sqrt{G(t-G)} + M_5[\alpha(G)]^{C_0/(C_0+2)}(t-G)\right) \\
&\quad + M_8 T^{\alpha_x - 1} \sum_{i=0}^{t-1} [\alpha_{mix}(i)]^{(C_0+1)/(C_0+2)}.
\end{aligned}$$

The first part of Proposition 8 shows that the big model that uses the near-unit root predictor is expected to outperform the smaller model that excludes this predictor and establishes bounds on the expected squared error gains from using the big model. Note that whenever $\alpha_x < 1/2$, the expected gain from using the big model grows without bounds as the sample size gets big.

To better understand the last two claims in Proposition 8, consider the simple case with $\alpha_{mix}(i) \leq \tau^i \ \forall i \geq 1$ for some constant $\tau \in (0, 1)$. Then

$$\min_{1\leq G\leq t} \left(M_3 G + M_4\sqrt{G(t-G)} + M_5[\alpha(G)]^{C_0/(C_0+2)}(t-G)\right) \leq K_1\sqrt{\log T}$$

and $\sum_{i=0}^{t-1} [\alpha_{mix}(i)]^{(C_0+1)/(C_0+2)} \leq K_2$, where $K_1, K_2 > 0$ are constants. Therefore, in this case, $\mathbb{P}(\mathbb{E}(\Delta L_{t+1} \mid z_{1t}) \leq 0)$ and $\left|\frac{\mathbb{E}\Delta L_{t+1}\mathbf{1}\{\mathbb{E}(\Delta L_{t+1}|z_{1t})>0\}}{\mathbb{E}\Delta L_{t+1}} - 1\right|$ are at most of order

$O(T^{-1}\sqrt{\log T} + T^{\alpha_x - 1})$.

Hence, if $\alpha_x < 1$, part (2) of Proposition 8 states that the probability of switching between the big and the small model goes to zero as the big model that accounts for the effect of the persistent regressor is expected to outperform the smaller (nested) model. Under the same conditions, Part (3) of Proposition 8 establishes that the relative performance of the switching rule and the big model is the same. This is unsurprising since, by part (2), there is not expected to be any switching between the big and the small models.

5.4 Choice of Monitoring Instruments

Our analysis is suggestive of which monitoring instruments to use in the switching rule. For example, suppose that $\mathbb{E}(\varepsilon_{t+1} \mid x_{1,t}, x_{2,t}) = 0$. Using the squared forecasts $z_{1t} = (\hat{y}^2_{1,t+1}, \hat{y}^2_{2,t+1})'$ as test instruments, we have $\mathbb{E}(\Delta L_{t+1} \mid z_{1t}) = \beta_2^2 x^2_{2,t} - \beta_1^2 x^2_{1,t} = \hat{y}^2_{2,t+1} - \hat{y}^2_{1,t+1}$ and so we can regress ΔL_{t+1} on $\hat{y}^2_{1,t+1}$ and $\hat{y}^2_{2,t+1}$ to obtain $\mathbb{E}(\Delta L_{t+1} \mid z_{1t})$.[23]

Alternatively, for regressors whose squared value is persistent, we can simply use the lagged value of ΔL_{t+1} as a monitoring instrument. This is an easy strategy since such lags are always observed even if $x_{1,t}$ and $x_{2,t}$ are not observable to the forecast evaluator, as in the case of survey forecasts. To see why this works, suppose that one of the regressors is persistent in squares, e.g., $x_{2,t} = \rho x_{2,t-1} + \sqrt{1-\rho^2} u_t$, where $\{x_{1,t}\}$, $\{u_t\}$ and $\{\varepsilon_{t+1}\}$ are mutually independent i.i.d sequences of standard normal random variables.[24] It is not hard to show that $Var(\Delta L_{t+1}) = 2(\beta_1^4 + \beta_2^4) + 4(\beta_1^2 + \beta_2^2)$ and $cov[\Delta L_{t+1}, \Delta L_t] = 2\rho^2 \beta_2^4$, so that, ignoring estimation error,

$$corr(\Delta L_{t+1}, \Delta L_t) = \frac{\rho^2 \beta_2^4}{(\beta_1^4 + \beta_2^4) + 2(\beta_1^2 + \beta_2^2)}.$$

Because the correlation between ΔL_{t+1} and ΔL_t is always nonnegative, a simple switching rule is to choose model 1 if and only if this model outperformed model 2 in the current period, i.e., if $\Delta L_t < 0$.[25]

[23] Even if $x_{j,t}$ is not observable, we always observe $\hat{y}_{j,t+1|t}$, functions of which which can therefore serve as z_{1t} in the switching rule.

[24] This situation might arise if one or both of the x variables has ARCH-type heteroskedasticity dynamics which introduces persistence in its squared value.

[25] In practice, we might need to take into account that $\mathbb{E}(\Delta L_{t+1}) \neq 0$ and thus a regression-based switching rule might be more appropriate.

5.5 Sources of Gains from Monitoring

Gains from monitoring forecasting performance arise from the non-zero correlation between $x_t\varepsilon_{t+1}$ and z_{1t}. It is worth briefly discussing how this non-zero correlation might arise. We consider three possible sources, namely time-varying parameters, model misspecification, and persistent estimation errors. Throughout the analysis we focus on the simple nested case in which the small model forecasts zero while the big model assumes a linear relation between y_{t+1} and x_t, i.e., i.e., $y_{t+1} = \beta x_t + \varepsilon_{t+1}$. Moreover, again we ignore estimation error.

5.5.1 Time-varying Parameters

Using the nested setup with a small and a big forecasting model, consider the case where the parameters of the data generating process are time-varying, i.e., $y_{t+1} = \beta_t x_t + \varepsilon_{t+1}$ follows a stationary but time-varying process, while the big forecasting model assumes constant parameters. In this case, the big model uses the "long-run mean", $\bar{\beta} = t^{-1}\sum_{\tau=1}^{t}\beta_\tau$, so that

$$\Delta L_{t+1} = (y_{t+1} - 0)^2 - \left(y_{t+1} - \bar{\beta}x_t\right)^2 = \bar{\beta}(2\beta_t - \bar{\beta})x_t^2 + 2\bar{\beta}x_t\varepsilon_{t+1}. \tag{25}$$

Suppose that x_t and z_{1t} are independent. Then

$$\mathbb{E}(\Delta L_{t+1} \mid z_{1t}) = \left[\bar{\beta}^2 + 2\bar{\beta}\mathbb{E}\left((\beta_t - \bar{\beta}) \mid z_{1t}\right)\right]\mathbb{E}x_t^2. \tag{26}$$

It is not hard to see that the switching rule based on z_{1t} is expected to outperform both the big and the small forecasting models as long as the following two events both have positive probability: $\bar{\beta}^2 + 2\bar{\beta}\mathbb{E}\left((\beta_t - \bar{\beta}) \mid z_{1t}\right) < 0$ and $\bar{\beta}^2 + 2\bar{\beta}\mathbb{E}\left((\beta_t - \bar{\beta}) \mid z_{1t}\right) > 0$.

5.5.2 Misspecified Forecasting Model

Suppose the data generating process takes the form

$$y_{t+1} = f(x_t) + \varepsilon_{t+1}, \tag{27}$$

where $f(x_t)$ is a general (nonlinear) function of x_t and we assume that $\mathbb{E}[x_t] = \mathbb{E}[\varepsilon_{t+1}] = \mathbb{E}[x_t\varepsilon_{t+1}] = 0$.

Under these assumptions, $\beta = \mathbb{E}[x_t y_{t+1}]/\mathbb{E}x_t^2 = \mathbb{E}(x_t f(x_t))/\mathbb{E}x_t^2$ and so the loss

differential is given by

$$\Delta L_{t+1} = (y_{t+1} - 0)^2 - (y_{t+1} - \beta x_t)^2 = \beta x_t(2f(x_t) - \beta x_t) + 2\beta x_t \varepsilon_{t+1}. \tag{28}$$

Again, it is easy to construct examples where the monitoring instrument z_{1t} is correlated with $x_t(2f(x_t) - \beta x_t)$ and so might be used to increase the predictive accuracy of either of the underlying forecasting models.

As a concrete example of how this might work, suppose that $x_t, z_{1t}, \xi_{t+1} \sim N(0,1)$ are independent normal random variables and the forecast error for the big model is given by $\varepsilon_{t+1} = \text{sign}(x_t z_{1t}) |\xi_{t+1}|$. Then we have $\mathbb{E}[x_t z_{1t}] = \mathbb{E}[x_t \varepsilon_{t+1}] = \mathbb{E}[z_{1t}\varepsilon_{t+1}] = 0$ and the covariance matrix of $(x_t, z_{1t}, \varepsilon_{t+1})$ is the 3×3 identity matrix. From (22), the expected loss differential is $\mathbb{E}[\Delta L_{t+1}] = \beta^2$. Although z_{1t} is independent of (x_t, ε_{t+1}), z_{1t} contains information on the mean of ΔL_{t+1}:

$$\mathbb{E}\left(\Delta L_{t+1} | z_{1t}\right) = \beta^2 + 2\beta \mathbb{E}\left(x_t \text{sign}(x_t z_{1t}) |\xi_{t+1}| \mid z_{1t}\right) = \begin{cases} \beta^2 + 2\beta a & \text{if } z_{1t} > 0 \\ \beta^2 & \text{if } z_{1t} = 0 \\ \beta^2 - 2\beta a & \text{if } z_{1t} < 0 \end{cases},$$

where $a = \mathbb{E}|x_t \xi_{t+1}| > 0$. For this case the sign of z_{1t} contains information about the loss differential and thus can be used to monitor the models' (relative) forecasting performance. For example, if $\beta^2 - 2\beta a < 0 < \beta^2 + 2\beta a$, then one should use the big model if and only if $z_{1t} > 0$.

6 Simulation Results

This section presents results from a set of Monte Carlo simulations which illustrate the theoretical analysis in sections 3 and 4. For the nested case we show the joint effects of varying the strength of the predictor and the monitoring instrument on the predictive performance of (i) a small forecasting model; (ii) a big forecasting model; and (iii) a switching rule. We also consider alternative forecasting methods based on augmenting the forecasting model with the monitoring instrument, a pre-test for determining whether to include a predictor, and an equal-weighted forecast combination.

For each point in time $t \geq m + n + 1$, define the rolling window estimator for a

model that includes x_{it} as a predictor

$$\hat{\beta}_{it} = \left[\sum_{s=t-n}^{t-1} x_{is}x_{is}\right]^{-1}\left[\sum_{s=t-n}^{t-1} x_{is}y_{s+1}\right].$$

with resulting forecast $\hat{y}_{i,t+1|t} = \hat{\beta}_{i,t}x_t$. The squared error loss differential of models 1 versus 2 is given by

$$\Delta L_{t+1} = (y_{t+1} - \hat{\beta}_{1,t}x_{1t})^2 - (y_{t+1} - \hat{\beta}_{2,t}x_{2t})^2. \tag{29}$$

To evaluate the switching rule, for $t \geq m+n+1$, define the estimates from regressing ΔL_{t+1} on $(1, z_{1t})$:

$$(\hat{\theta}_{0,t}, \hat{\theta}_{1,t})' = \left[\sum_{s=t-m}^{t-1} (1, z_{1s})(1, z_{1s})'\right]^{-1}\left[\sum_{s=t-m}^{t-1} (1, z_{1s})'\Delta L_{s+1}\right]$$

and the associated conditional forecast of the loss differential

$$\widehat{\Delta L}_{t+1|t} = \hat{\theta}_{0,t} + \hat{\theta}'_{1,t}z_t. \tag{30}$$

Forecasts from the switching rule take the form

$$\hat{y}_{SW,t+1|t} = \mathbf{1}\{\widehat{\Delta L}_{t+1|t} \leq 0\}\hat{\beta}_{1,t}x_{1t} + \mathbf{1}\{\widehat{\Delta L}_{t+1|t} > 0\}\hat{\beta}_{2,t}x_{2t}$$

For each simulated sample we compute the mean squared errors of the two forecasts and for the switching rule as $MS\mathbb{E}^j = p^{-1}\sum_{t=m+n+1}^{T-1}(y_{t+1} - x_{jt}\hat{\beta}_{j,t})^2$, and $MS\mathbb{E}^{SW} = p^{-1}\sum_{t=m+n+1}^{T-1}(y_{t+1} - \hat{y}_{SW,t+1|t})^2$, where $T = n + m + p$.

To shed light on the power of the GW test for equal forecasting performance, for each sample we compute ΔL_{t+1} and $\widehat{\Delta L}_{t+1}$ from (29) and (30), respectively. Statistical significance of the loss differentials of forecasts $j_1$and j_2 is then evaluated using the test statistics

$$J_T^{j_1-j_2} = p^{-1/2}\sum_{t=m+n+1}^{T-1}\frac{\Delta L_{t+1}^{j_1-j_2}}{\hat{\sigma}(\{\Delta L_{t+1}^{j_1-j_2}\})}, \tag{31}$$

where $\hat{\sigma}(\{\Delta L_{t+1}^{j_1-j_2}\})$ is the Newey-West standard deviation of the loss differential

between models j_1 and j_2.

6.1 Nested case

In the nested case, data are generated from a simple linear regression model

$$y_{t+1} = \beta x_t + \varepsilon_{t+1}, \tag{32}$$

where $x_t \sim i.i.d.U(-1,1)$. The residual ε_{t+1} is generated as follows. Let $s_{t+1} \in \{0,1\}$ be a binary random variable such that $\mathbb{P}(s_{t+1} = 1 \mid x_t > 0) = \mu + \delta$ and $\mathbb{P}(s_{t+1} = 1 \mid x_t \leq 0) = \mu - \delta$, where $\mu = 1/2$. Define

$$\varepsilon_{t+1} = s_{t+1}Q_{1,t} + (1 - s_{t+1})Q_{2,t}, \tag{33}$$

where $Q_{1,t}$ and $Q_{2,t}$ are $N(0,1)$ random variables that are mutually independent and independent of s_{t+1} and x_t. To control the correlation between the residual in (32) and the monitoring instrument, z_{1t}, we generate the latter as

$$z_{1t} = a_1 Q_{1,t} + a_2 Q_{2,t}, \tag{34}$$

where $a_1 = 1$ and $a_2 = -1$. It is now easy to see that $\mathbb{E}x_t\varepsilon_{t+1} = \mathbb{E}z_{1t}\varepsilon_{t+1} = \mathbb{E}\varepsilon_{t+1} = \mathbb{E}z_{1t} = \mathbb{E}x_t = 0$ and

$$Corr(x_t\varepsilon_{t+1}, z_{1t}) = \sqrt{\frac{3}{2}}\delta.$$

Our simulations set $\beta = 3n^{-\alpha_x}$ and we choose δ such that $Corr(x_t\varepsilon_{t+1}, z_{1t}) = 0.6n^{-\alpha_z}$. We report the outcome of 5,000 simulations based on a sample size $(n, m, p) = (100, 100, 200)$, so that $T = 400$.

Table 1 presents results from the simulations. In each panel we show the proportion of simulations for which the null of equal predictive accuracy is rejected against a one-sided alternative, using the J_T test in (31) and a 5% size. The higher the value of $\mathbb{P}(J_T^{1-2} > 1.64)$, the stronger the evidence that forecasts from model 1 generate larger out-of-sample MSE values than forecasts from model 2.

First consider the performance of the big versus the small forecasting model (top row) in Panel A. When $\alpha_x \leq 0.25$, the big model produces far more accurate forecasts than the small model–as indicated by rejection rates exceeding 99% for a test that the MSE of the small model exceeds that of the big model. As α_x rises to 0.5, the

rejection rate declines to around 10%, and the forecasting performances of the big and small models become increasingly similar. In fact, for $\alpha_x = 0.75$ and $\alpha_x = 1$, the small model produces lower MSE values than the big model, consistent with our theoretical analysis (Proposition 4).

Turning to the comparison of the switching rule and the small forecasting model (Panel B), we see that the switching rule strongly dominates the small model (with rejection rates above 85%) if either (i) the predictor is strong, i.e., $\alpha_x \leq 0.25$, or the monitoring instrument is strong ($\alpha_z \leq 0.1$) and the predictor is not weaker than local-to-zero ($\alpha_x < 0.5$). In the first case, the accuracy of the monitoring instrument does not matter because one option is to always select the big model which, on average, performs better than the small model provided that α_x is small, ensuring that the predictor is informative. The better average performance of the big model relative to the small model is picked up by the intercept θ_0 in the switching rule and so holds independently of the value of θ_1. In the second case, a precise monitoring instrument allows for accurate determination of when the small or the big forecasting model is likely to be best–even in situations with a less accurate forecasting signal. As both α_x and α_z rise beyond these values, the predictive accuracy of the switching rule, measured relative to the small model, deteriorates, consistent with what we would expect from the theoretical analysis (Proposition 5).

Measured relative to the big model, Panel C of Table 1 shows that the predictive accuracy of the switching rule is, in contrast, very poor when the predictive instrument (x_t) is quite accurate, i.e., when α_x is small. However, as α_x rises above 0.5, we start to see that the switching rule increasingly dominates the big forecasting model. Interestingly, even when α_x is small and so the predictor is accurate, the switching rule can produce more accurate forecasts than the big model provided that the monitoring instrument is very accurate, i.e., $\alpha_z \leq 0.2$.

Table 2 reports the performance of the switching rule measured relative to three alternative, widely used, forecasting methods. Panel A shows that the switching rule dominates an equal-weighted average of the two forecasting models if either (i) the big model is very good (i.e., α_x is small), regardless of the accuracy of the monitoring instrument; or (ii) if the monitoring instrument is very accurate (small α_z), regardless of the precision of the predictor. The switching rule only loses out to the equal-weighted forecast combination if both the monitoring instrument and the predictor are poor, i.e., if α_x and α_z are both large. Panel B compares the performance of the

switching rule to that of a pre-test approach that includes x in the forecasting model only if its slope coefficient is statistically significant using a t-test. We find that the switching rule performs best provided that the predictor is not too accurate and the monitoring instrument is fairly accurate (small α_z). Panel C shows that the switching rule is particularly likely to generate more accurate forecasts than those from an augmented model that includes the monitoring instrument, z_{1t}, as a predictor in the big forecasting model provided that the monitoring instrument is accurate (α_z is low) and the predictor variable is not too strong. Note the non-monotonic pattern in the rejection rates which first increase, then decline as a function of α_x (for small α_z) or conversely first decline, then rise (for large α_z) when we measure the performance of the switching rule relative to the pre-test or augmentation methods.

6.2 Non-nested case

For the non-nested case, let $\{(x_{1,t}, z_{1t,1}, s_{1,t}, \varepsilon_{1,t+1})\}_{t=1}^{T}$ and $\{(x_{2,t}, z_{1t,2}, s_{2,t}, \varepsilon_{2,t+1})\}_{t=1}^{T}$ be independent copies of the process $\{(x_t, z_{1t}, s_t, \varepsilon_{t+1})\}_{t=1}^{T}$ in (32) - (34) such that in generating $\{(x_{j,t}, z_{1t,j}, s_{j,t}, \varepsilon_{j,t+1})\}_{t=1}^{T}$, we use $\delta_j = \sqrt{2/3} \times 0.6n^{-\alpha_{z,j}}$ for $j \in \{1, 2\}$. Then we set $z_{1t} = (z_{1t,1} + z_{1t,2})/2$, $\varepsilon_{t+1} = (\varepsilon_{1,t+1} + \varepsilon_{2,t+1})/2$ and $\beta_j = 3n^{-\alpha_{x,j}}$ in

$$y_{t+1} = \beta_1 x_{1,t} + \beta_2 x_{2,t} + \varepsilon_{t+1}. \tag{35}$$

Table 3 shows the outcome of three comparisons of predictive accuracy for model 1 versus model 2, model 1 against the switching rule, and model 2 versus the switching rule. We let $\alpha_{x,1}$ and $\alpha_{x,2}$ take values of $\{0, 0.25, 0.5, 1\}$. The four panels in the table correspond to different combinations of the accuracy of the monitoring instrument for models 1 and 2, with $\{\alpha_{z,1}, \alpha_{z,2}\} = \{0, 0\}$ (panel A), $\{\alpha_{z,1}, \alpha_{z,2}\} = \{0, 1\}$ (panel B), $\{\alpha_{z,1}, \alpha_{z,2}\} = \{0.5, 0.5\}$ (panel C), and $\{\alpha_{z,1}, \alpha_{z,2}\} = \{1, 1\}$ (panel D). When $\alpha_{z,1} = 0$, we find that the first model produces significantly more accurate forecasts than the second model (rejection rates near zero) provided that $\alpha_{x,2}$ is larger than $\alpha_{x,1}$, so that the predictor for model 1 is more accurate than that for model 2, and $x_{1,t}$ is a reasonably strong predictor ($\alpha_{x,1} \leq 0.25$). Conversely, by symmetry, the second model produces significantly more accurate forecasts than the first model (rejection rates near one) when $\alpha_{x,2} \leq 0.25$ and $\alpha_{x,2}$ is smaller than $\alpha_{x,1}$, so that the predictor for model 2 is quite strong and more accurate than that used by the first model.

Next consider the performance of the switching rule relative to the individual

forecasts when both monitoring instruments are strong ($\alpha_{z,1} = \alpha_{z,2} = 0$ in panel A). The switching approach generates better forecasts than model 1 when $\alpha_{x,1} \geq 0.5$ so that this model uses a fairly weak predictor and switching away from this is beneficial. Interestingly, this holds even when $\alpha_{x,1}$ is quite large, so that the predictor used by the second forecasting model is quite poor. The switching rule performs distinctly worse relative to always using model 1 if $\alpha_{x,1}$ is very small, while $\alpha_{x,2}$ is high, so that there is little reason for switching away from model 1 and towards model 2. The opposite pattern holds when we compare the performance of the switching rule to that of model 2. Relative to model 2, the switching rule generates more accurate forecasts for small values of $\alpha_{x,1}$ and big values of $\alpha_{x,2}$ while it underperforms for the reverse scenario. Results only change marginally when we use a weaker instrument for model 2, i.e., when $\alpha_{z,2} = 1$ (Panel B). Interestingly, when both predictor variables are local-to-zero, i.e., $\alpha_{x,1} = \alpha_{x,2} = 1/2$, there is more than a 25% chance that the switching rule produces more accurate than both model 1 and model 2.

Panels A and B assume that at least one monitoring instrument ($\alpha_{z,1}$) is highly accurate. In Panels C and D, we instead let both instruments be relatively poor. The absence of accurate monitoring instruments means that the probability that the switching rule outperforms both of the underlying forecasting models deteriorates significantly. In particular, the probabilities that the switching rule will perform better than the underlying models when the signals are neither very weak nor very strong, i.e., $\alpha_{x,i}$ is at 0.25 or 0.50, are now much reduced compared to the case where at least one of the monitoring instruments is strong.

7 Empirical Analysis

We next provide empirical illustrations of how monitoring instruments can be used, first, to understand time variation and state dependencies in the performance of prediction models and, second, to evaluate the performance of the switching rule. We present two empirical applications. Our first application is to inflation forecasting. To illustrate the non-nested case, we compare the Federal Reserve's Greenbook forecasts to forecasts from the Survey of Professional forecasters. For the nested case, we also compare the accuracy of a simple backward-looking Phillips curve model to that of an autoregressive specification. This example displays stronger evidence of predictability and is interesting given the strong evidence of parameter instability found

for inflation forecasting models; see, e.g., Stock and Watson (2003). Second, we study predictability of U.S. monthly stock market returns–a case with weak predictability and large estimation errors.

7.1 Inflation Forecasts

Our first empirical application looks at the predictability of U.S. inflation. Inflation forecasting has generated a lot of interest in the economics literature despite appearing to have become more difficult over time (Stock and Watson (2007)). We perform our analysis in two stages. First, we compare forecasts from the Federal Reserve Greenbook to forecasts from the Survey of Professional Forecasters (SPF)-the data underlying the plots in Figure 1. Second, we compute recursively generated forecasts from a backward-looking Phillips curve and compare them to forecasts from a simple autoregressive model.

7.1.1 Greenbook versus SPF Forecasts

Our first application compares the predictive accuracy of the Federal Reserve's quarterly Greenbook forecasts of the GDP price deflator to the mean forecast of the same variable from the SPF using forecast horizons ranging from one through four quarters over the sample period 1968Q4-2010Q4.[26] It is highly unlikely that the two sets of forecasts use the exact same predictor variables, so this application represents the non-nested case.

The first column in Table 4 shows t-statistics from a simple Diebold-Mariano regression of the squared error loss associated with the Greenbook forecasts minus that from the SPF on an intercept. Negative values suggest that the Greenbook forecasts, on average, are more accurate than the SPF forecasts and this is indeed what we find. The subsequent columns show Diebold-Mariano t-statistics from comparing the MSE performance of the switching rule versus always using the Greenbook forecasts (labeled GB-SW) or comparing the conditional selection rule versus always using the SPF forecasts (labeled SPF-SW). In these comparisons, positive values indicate that the conditional switching rule performs better than the individual forecasts. We use four different test instruments, namely the unemployment gap (UG) proposed by

[26]Data on the forecasts are obtained from the Federal Reserve Bank of Philadelphia.

Stock and Watson (2010),[27] a measure of macroeconomic uncertainty proposed by Jurado et al. (2015) (labeled $\mathcal{U}$ and constructed as the first principal component from a variety of uncertainty measures), the lagged loss measured over the most recent four quarters, $\Delta\bar{L}_{t-3:t} = (1/4)\sum_{\tau=1}^{4}\Delta L_{t+1-\tau}$, along with the squared difference $\hat{y}_{1t}^2 - \hat{y}_{2t}^2$, again averaged over the most recent four quarters, i.e., $\Delta\hat{y}_{t-3:t}^2 = (1/4)\sum_{\tau=0}^{3}(\hat{y}_{1t+1-\tau}^2 - \hat{y}_{2t+1-\tau}^2)$. We consider both a 10-year rolling estimation window for the switching regression ($m = 40$, panels A and C) and a 15-year rolling estimation window ($m = 60$, panels B and D) and report results that measure the actual inflation figure using either the final revision to the GDP deflator (panels A and B) or real-time vintage estimates (panels C and D).

Overall, the results show that it is easy to find monitoring instruments that allow us to identify periods where the Greenbook forecasts outperform the SPF forecasts. This is not surprising because we know that the Greenbook forecasts are better on average than the SPF forecasts (column 1) and so a conditional selection rule that always prefers the Greenbook forecasts should also produce good results. More interestingly, it is also possible to find instruments for which the conditional switching rule improves significantly on the Greenbook forecasts for at least some horizons.

As an illustration of how our approach can be used to monitor forecasting performance at a higher frequency than the daily data, we regressed the quarterly series of squared error differences (Greenbook minus SPF) on the term spread measured at the end of the previous quarter.[28] We use this monitoring instrument in part because it is easy to construct on a daily basis. While the dependent variable (the squared error loss difference) is only observed once a quarter, in this case the predictor, z_{1t}, is observed daily and so, using our estimates of θ, we can generate daily estimates of the conditionally expected loss, ΔL_{t+1}. These estimates are plotted in Figure 3 over the sample period 2000-2012. We see that the expected loss difference fluctuates substantially and reverses sign during the financial crisis. This suggests that whereas the Greenbook forecasts were expected to be less accurate leading up to the financial

[27] This is measured as $z_{1t} = ug_t = u_t - \min(u_t, ..., u_{t-35})$, where u_t is the unemployment rate in month t so the gap is computed relative to the unemployment rate during the previous three years. This variable rises during recessions and comes down during the early expansion phase of most economic cycles and can be viewed as a "real time" alternative to the NBER recession indicator which is only released with several months' lag.

[28] The term spread is the difference in the yield of a long (10-year) Treasury bond and the yield on a short (3-month) T-bill. The estimate of θ_1 is statistically significant in the monitoring regression which uses a forecast horizon of $h = 2$ quarters.

crisis, they were expected to perform better than the SPF forecasts during the crisis.

This application shows that our method for monitoring forecasting performance can be beneficial for decision makers such as central banks in "benchmarking" their forecasts against forecasts from other sources to see if there is evidence in real time that the accuracy of their forecasts is lagging behind competing forecasts. While it may not be feasible for such decision makers to adopt another agency's forecasts, evidence of inferior performance would suggest the need to improve on their own forecasting methods.

7.1.2 Forecasts from a Backward looking Phillips Curve Model

Our second inflation example compares the forecasting performance of a backward-looking Phillips curve to that of a nested autoregressive model. First, define the annualized quarterly inflation rate as $\pi_t = 400 \times \log(\mathbb{P}_t/\mathbb{P}_{t-1})$, where $\mathbb{P}_t$ is the quarterly price index for the U.S. GDP deflator. Because the quarterly inflation rate is highly persistent, we model the change in the inflation rate, $\Delta\pi_t$. The small model is assumed to be an AR(4) specification:

$$\Delta\pi_{t+1} = \beta_0 + \sum_{i=1}^{4} \beta_i \Delta\pi_{t+1-i} + \varepsilon_{1t+1}. \tag{36}$$

Autoregressive models such as (36) have proven difficult to outperform in many forecast comparisons. The big forecasting model (B) adds the lagged unemployment rate, u_t, to (36) to obtain a backward-looking Phillips curve model of the form

$$\Delta\pi_{t+1} = \beta_0 + \sum_{i=1}^{4} \beta_i \Delta\pi_{t+1-i} + \gamma_1 u_t + \varepsilon_{2t+1}, \tag{37}$$

As our monitoring instrument, z_{1t}, we use the first principal component ($\mathbb{P}C_t$) extracted from a large cross-section of more than 100 macroeconomic variables, using the data set provided by Jurado et al. (2015).

Our sample runs from 1950Q1 to 2011Q4 and we use an estimation window of 50 quarterly observations. The monitoring window used to implement the switching rule is also set at 50 quarterly observations ($n = m = 50$). Accounting for lags in the model, this leaves 103 observations over the period 1986Q2-2011Q4 for out-of-sample forecast evaluation.

The upper panel in Figure 4 compares out-of-sample forecasts of $\Delta L_{t+1} = e^2_{S,t+1} - e^2_{B,t+1}$, generated by regressing ΔL_{t+1} on a constant and $z_{1t} = PC_t$, to the realized values of ΔL_{t+1}. For most of the sample, the expected loss $\widehat{\Delta L}_{t+1|t}$ hovers around zero with frequent shifts in which model is expected to perform best. Even so, the baseline AR(4) model is expected to produce the most accurate forecasts for long stretches of time during the intervals 1988-1997, 2006-2008 and 2010-2011. The periods where the backward-looking Phillips curve (37) is expected to perform best are typically shorter-lived with exception of the earliest part of the sample 1986-1987.

The lower panel in Figure 4 shows the cumulative differences in MSE values for four different model comparisons. Gray areas show periods where the small model is expected to generate a smaller loss than the big model. The blue line tracks the performance of the big model (37) against the small model (36). Adding the unemployment rate to the AR specification does not change the precision of the forecasts by much until the second half of 2009 at which point it leads this model to greatly underperform the smaller model. This is related to the extreme value of the realized inflation rate in the fall of 2009 and causes the big model to marginally underperform the small model with a test statistic $J_T^{S-B} = -0.08$ for the full sample.

Comparing an augmented model that adds z_{1t} as an additional predictor to the small model in (36),

$$\Delta\pi_{t+1} = \beta_0 + \sum_{i=1}^{4} \beta_i \Delta\pi_{t+1-i} + \gamma_1 u_t + \gamma_2 \mathbb{P}C_t + \varepsilon_{t+1}, \tag{38}$$

the augmented model (38) performs worse than the big model (37) although the difference (t-statistic of 0.92) is not statistically significant.

The conditional switching rule that chooses the forecasting model based on the expected value of the forecast differential ΔL_{t+1} produces more accurate forecasts than using either of these models. The red line shows that the conditional switching rule produces better forecasts than the autoregressive model for much of the sample up to the fall of 2009 where, again, this approach gets on the wrong side of the extreme value of the inflation rate and so underperforms the small model. On average, across the full out-of-sample period, the conditional switching rule performs marginally better than the small model. Compared to the big forecasting model (black line), the switching rule outperforms most of the time. Moreover, the difference in MSE value is statistically significant with a t-statistic of 2.15.

We conclude the following from these results. First, always using the unemployment rate to predict inflation as is done by the big model (37) does not lead to improved forecasts. However, it seems that there are states where this variable adds predictive ability to a simple autoregressive model of inflation, the fall of 2009 excluded, and these states can be predicted by means of a variable that summarizes aggregate economic information. Second, consistent with our earlier theoretical analysis, using the monitoring instrument z_{1t} to switch between the small and big models leads to significantly better forecasting performance than simply using this instrument as a predictor variable in the conventional way as is done in equation (38).

7.2 Predictability of US Stock Market Returns

Our second application considers predictability of U.S. stock market returns. Specifically, our dependent variable is the monthly excess return on the S&P500 index, measured net of a short T-bill rate. Welch and Goyal (2008) analyze a large set of univariate prediction models and find that none of these is capable of generating smaller out-of-sample MSE values than a simple prevailing mean (constant equity premium) model. We follow their analysis in using the prevailing mean model as our benchmark (small model):

$$y_{t+1} = \mu + \varepsilon_{St+1}. \tag{39}$$

This is compared to a univariate prediction model that uses the lagged value of the one-month T-bill rate as a predictor and thus takes the form

$$y_{t+1} = \mu + \beta x_t + \varepsilon_{Bt+1}. \tag{40}$$

In addition, following Rapach et al. (2010), we consider a simple equal-weighted forecast combination scheme that combines forecasts from 14 univariate models. This is shown by Rapach et al. (2010) to produce more accurate forecasts than the individual univariate forecasting models whose performances are strongly affected by estimation error.[29] Forecasts from the equal weighted (EW) combination are computed as

$$\tilde{y}_{t+1|t} = \frac{1}{14}\sum_{i=1}^{14} \hat{y}_{i,t+1|t}, \tag{41}$$

[29]Such combined forecasts also perform far better than forecasts from a multivariate "kitchen sink" regression that includes all 14 predictor variables.

where $\hat{y}_{i,t+1|t} = \hat{\mu}_i + \hat{\beta}_i x_{i,t}$ is the forecast from the ith univariate return prediction model.

All data are obtained from Goyal's web site and cover the sample 1927-2013. We use a 20-year rolling window to estimate the parameters of the underlying forecasting models ($n = 240$) and also use 20 years of out-of-sample forecasts to conduct switching rule regressions ($m = 240$) and compute the expected loss differential in (6). Our analysis of the small and big models' out-of-sample forecasting performance thus runs from 1967 through 2013, a total of 47 years or 564 monthly observations. As monitoring instruments, we use a list similar to those used in the analysis of the inflation data

7.2.1 Empirical Findings

Table 5 reports the outcome of different tests of $H_0 : \mathbb{E}[\Delta L_{t+1}|Z_t] = 0$ versus the alternative $H_1 : \mathbb{E}[\Delta L_{t+1}|Z_t] \neq 0$ under squared error (MSE) loss. Panel A shows results for the equal-weighted forecast combination, while the subsequent panels show results from the univariate forecasting model that uses one of our three predictor variables. In each panel, the top row reports results for $Z_t = 1$ corresponding to a test of equal unconditional expected forecasting performance, i.e., a conventional Diebold and Mariano (1995) test of $H_0 : \mathbb{E}[\Delta L_{t+1}] = 0$. .

The subsequent rows present estimates from a linear regression of the loss differential on an intercept and the monitoring instrument, $\Delta L_{t+1} = \theta_0 + \theta_1 z_{1t} + u_{t+1}$. The first two columns show t-statistics for the estimates of the associated regression coefficients, θ_0, θ_1 followed by $100 \times R^2$ of this regression in the third column. The R^2 value shows how much of the difference in the big and small models' squared forecast errors can be predicted by the various monitoring instruments. The fourth column, labeled GW, shows the p-value of the Giacomini-White test (7) based on this regression. Low p-values indicate rejection of the null of equal conditional forecasting performance and thus suggest that differences in the performance of the big versus small forecasting models can be tracked by means of information in Z_t.

First consider the results for the forecast combination (Panel A). With a t-statistic of 1.82, there is some evidence that the forecast combination produces a lower MSE "on average" than the benchmark prevailing mean model. Turning to the conditional performance estimates, the GW test rejects the null of equal predictive performance for three of four instruments at the 5% critical level or better. Some of this is driven

by the significance of the intercept, i.e., the better average performance of the forecast combination compared to the performance of the benchmark, prevailing mean model. However, the lagged value of the squared forecast differential is significantly negatively correlated with the future loss differential, $\triangle L_{t+1}$.

Turning to the results for the univariate forecasts, we find that the T-bill rate model's average MSE forecasting performance is actually worse than that of the prevailing mean, as indicated by the negative estimate of θ_0 in the first row of Panel B. Interestingly, when used as monitoring instruments, both the unemployment gap and the squared forecast difference are capable of identifying time variation in the loss differential $\triangle L_{t+1}$ of the prevailing mean model relative to the T-bill model, resulting in a rejection of the GW test of equal predictive accuracy at the 5% level for the former instrument.

Figure 5 provides further details of this result using the prevailing mean model as the small model and comparing its performance to a model that adds the T-bill rate (big model) and uses the unemployment gap variable as a monitoring instrument. The lines in the figure show the cumulative sum of squared forecast error differentials for the small minus the big model (blue line), the small model minus the switching rule (red), the big model minus the switching rule (black) and the "biggest" model augmented with the monitoring instrument (the unemployment gap) versus the small model (purple). When any of these graphs is rising and positive, the first model underperforms (produces higher squared errors) the second model and vice versa.[30] From these graphs it follows that the big model produces large gains in predictive accuracy (measured relative to the small model) between 1968 and 1975 only to see these gains disappear between 1975 and 1981 and, again, after 2009. Overall, measured across the full sample, adding the T-bill rate to the constant expected return model does not lead to lower cumulative squared error values. In contrast, the switching rule produces more accurate forecasts than both the small (red line) and big (black line) prediction models. Relative to the big model that always includes the T-bill rate as a predictor, the switching rule avoids the very large deterioration in forecasting performance observed in 1975-1976 and after 2009. Notably, the switching rule takes advantage of the T-bill rate model's good performance in the early seventies

[30]During periods (marked in gray) where the small model is expected to generate more accurate forecasts than the big model, the switching rule chooses the former and so the cumulative loss differential line for the switching rule relative to the small model becomes flat.

and, again, during the early part of the global financial crisis. Notice, finally, that the biggest model that adds the monitoring instrument (here the unemployment gap) as a predictor underperforms the small model and falls far short of the performance of the switching approach that uses the same information to choose between the big and the small models.

8 Conclusion

We construct finite-sample bounds on the expected mean squared error performance of different forecasting models which account for parameter estimation error and hold both on average (unconditionally) or conditional on a set of monitoring instruments. Our analysis covers both the case with nested and non-nested forecasting models. We show that the possibility of establishing gains from monitoring the loss difference between competing forecasting models and selecting at each point in time the model with the smallest expected loss requires conditions on the accuracy of both the predictors used by the underlying forecasting models as well as the strength of the monitoring instruments.

Through Monte Carlo simulations we demonstrate that there is indeed scope for the presence of monitoring instruments to help improve forecasting performance. For a switching rule to work, at least one of the models must use predictors that are not too weak. None of the underlying forecasting models can be too dominant as, otherwise, there is little space for improvements by alternating between the two forecasting models. Cases with forecasting models that have broadly similar predictive accuracy are frequently encountered in empirical studies and so this situation seems to match commonly found situations.

Our empirical applications to inflation forecasting and predictability of stock market returns demonstrate that it is not difficult to find examples of monitoring instruments that do not have predictive power if added directly to a forecasting model but that the monitoring instruments nevertheless can add value by containing information on when a particular forecasting model is likely to perform best conditional on such information.

Appendix

This appendix provides proofs of the theoretical results in the paper. The appendix is structured as follows. Appendix A provides technical tools used in the proofs while the theoretical results in the main text are proved in Appendix B.

First, some comments on notation. Throughout the appendix, the constants do not depend on T, n or t. For a vector $x = (x_1, \cdots, x_p)' \in \mathbb{R}^p$, $\|x\|_r = (\sum_{i=1}^p |x_i|^r)^{1/r}$. For a random variable or vector X, let $\|X\|_{L^r(\mathbb{P})} = (\mathbb{E}\|X\|_r^r)^{1/r}$. For two sequences $a_T, b_T > 0$, we say that $a_T \asymp b_T$ if $a_T = O(b_T)$ and $b_T = O(a_T)$. For any real number $x \geq 0$, we define $\lfloor x \rfloor$ to be the largest integer no larger than x.

A Technical results used in the proofs

Lemma 1. *Let $\mathcal{F}$ and $\mathcal{G}$ be σ-algebras with strong mixing coefficient α. Let $X \in \mathcal{F}$ and $Y \in \mathcal{G}$ be random variables with $\mathbb{E}X = 0$. Suppose that $\|X\|_{L^p(\mathbb{P})} \leq C_1$, $\|Y\|_{L^q(\mathbb{P})} \leq C_2$ for some constants $C_1, C_2 > 0$ and $p, q \in (1, \infty]$ satisfying $1/p + 1/q < 1$. Then $\mathbb{E}\,|\mathbb{E}(X|\mathcal{G})Y| \leq 8\alpha^{1-1/p-1/q} C_1 C_2$.*

Proof. Let $Z = \mathbb{E}(X \mid \mathcal{G})$. Define h to be the sign of YZ, i.e., $h = \mathbf{1}\{YZ > 0\} - \mathbf{1}\{YZ < 0\}$. Therefore, $Z, h \in \mathcal{G}$. We notice that

$$\mathbb{E}|ZY| \overset{(i)}{=} \mathbb{E}(ZYh) = \mathbb{E}\left[\mathbb{E}(X \mid \mathcal{G})Yh\right] \overset{(ii)}{=} \mathbb{E}(XYh) \overset{(iii)}{\leq} 8\alpha^{1-1/p-1/q} \|X\|_{L^p(\mathbb{P})} \|hY\|_{L^q(\mathbb{P})},$$

where (i) holds by $|ZY| = ZYh$, (ii) follows by $Y, h \in \mathcal{G}$ and the law of iterated expectations and (iii) holds by $\mathbb{E}X = 0$ and Davydov's Theorem (Theorem 3.7 of Bradley (2007)). By the above display and $\mathbb{P}(|h| \leq 1) = 1$, the desired result follows. □

Lemma 2. *Let X and Y satisfy that $\mathbb{E}|X|^{c_1}, \mathbb{E}|Y|^{c_2} \leq D$. Then $\mathbb{E}|XY|^v \leq D$, where $v = c_1 c_2/(c_1 + c_2)$.*

Proof. Let $p = c_1/v$ and $q = c_2/v$. Then $p^{-1} + q^{-1} = 1$. The result follows by Holder's inequality:

$$\mathbb{E}|XY|^v \leq \left(\mathbb{E}|X|^{vp}\right)^{1/p} \left(\mathbb{E}|Y|^{vq}\right)^{1/q} = \left(\mathbb{E}|X|^{c_1}\right)^{1/p} \left(\mathbb{E}|Y|^{c_2}\right)^{1/q} \leq D.$$

□

Lemma 3. *Let $\{X_i\}_{i=1}^n$ be independent random variables. Suppose that $\mathbb{E}X_i = 0$ and $\max_{1\le i\le n}\mathbb{E}|X_i|^p < K$ for some constants $p > 2$ and $K < \infty$. Then $\forall a, t > 0$,*

$$\mathbb{P}\left(\left|\sum_{i=1}^n X_i \mathbf{1}\{|X_i| \le a\}\right| \ge \sqrt{n}t + na^{1-p}K\right) \le 2\exp\left[-\frac{t^2}{2\left(a^{2-p}K + atn^{-1/2} + n^{-1}\sum_{i=1}^n \mathbb{E}X_i^2\right)}\right].$$

Proof. Let $\tilde{X}_i = X_i\mathbf{1}\{|X_i| \le a\}$, $Z_i = \tilde{X}_i - \mathbb{E}\tilde{X}_i$ and $B_n^2 = \sum_{i=1}^n \mathbb{E}Z_i^2$. Thus, $\mathbb{E}Z_i = 0$ and $\mathbb{P}(Z_i \le 2a) = 1$. It follows by Theorem 2.17 of Peña et al. (2008) that $\forall z > 0$

$$\mathbb{P}\left(\sum_{i=1}^n Z_i \ge z\right) \le \exp\left(-\frac{z^2}{2(B_n^2 + 2az)}\right).$$

Applying the same result to $\{-Z_i\}_{i=1}^n$, we obtain

$$\mathbb{P}\left(-\sum_{i=1}^n Z_i \ge z\right) \le \exp\left(-\frac{z^2}{2(B_n^2 + 2az)}\right).$$

The above two inequalities imply that

$$\mathbb{P}\left(\left|\sum_{i=1}^n Z_i\right| \ge z\right) \le 2\exp\left(-\frac{z^2}{2(B_n^2 + 2az)}\right). \tag{42}$$

We now bound $\mathbb{E}\tilde{X}_i$. Notice that

$$\begin{aligned}|\mathbb{E}\tilde{X}_i| = |\mathbb{E}(X_i - X_i\mathbf{1}\{|X_i| > a\})| &\overset{(i)}{=} |\mathbb{E}X_i\mathbf{1}\{|X_i| > a\}| \le \mathbb{E}|X_i|\mathbf{1}\{|X_i| > a\} \\ &\le \mathbb{E}\left|\frac{|X_i|^{p-1}}{a^{p-1}}\left(|X_i|\mathbf{1}\{|X_i| > a\}\right)\right| \le \mathbb{E}|X_i|^p a^{1-p},\end{aligned}$$

where (i) holds by $\mathbb{E}X_i = 0$. Therefore,

$$\left|\sum_{i=1}^n \mathbb{E}\tilde{X}_i\right| \le \sum_{i=1}^n |\mathbb{E}\tilde{X}_i| \le nKa^{1-p}. \tag{43}$$

Moreover, $\forall t > 0$,

$$\mathbb{P}\left(\left|\sum_{i=1}^n X_i\mathbf{1}\{|X_i| \le a\}\right| \ge \sqrt{n}t + na^{1-p}K\right)$$

$$
\begin{aligned}
&= \mathbb{P}\left(\left|\left(\sum_{i=1}^n Z_i\right) + \left(\sum_{i=1}^n \mathbb{E}\tilde{X}_i\right)\right| \geq \sqrt{n}t + na^{1-p}K\right) \\
&\leq \mathbb{P}\left(\left|\sum_{i=1}^n Z_i\right| + \left|\sum_{i=1}^n \mathbb{E}\tilde{X}_i\right| \geq \sqrt{n}t + na^{1-p}K\right) \\
&\overset{(i)}{\leq} \mathbb{P}\left(\left|\sum_{i=1}^n Z_i\right| \geq \sqrt{n}t\right) \overset{(ii)}{\leq} 2\exp\left(-\frac{nt^2}{2(B_n^2 + 2a\sqrt{n}t)}\right), \qquad (44)
\end{aligned}
$$

where (i) follows by (43) and (ii) follows by (42) with $z = \sqrt{n}t$.

It remains to bound B_n^2. Notice that

$$
\begin{aligned}
B_n^2 - \sum_{i=1}^n \mathbb{E}X_i^2 &= \sum_{i=1}^n \left[\mathbb{E}\tilde{X}_i^2 - \mathbb{E}X_i^2 - (\mathbb{E}\tilde{X}_i)^2\right] \\
&= \sum_{i=1}^n \left[\mathbb{E}X_i^2 \mathbf{1}\{|X_i| > a\} - (\mathbb{E}\tilde{X}_i)^2\right] \\
&\leq \sum_{i=1}^n \mathbb{E}X_i^2 \mathbf{1}\{|X_i| > a\} \\
&\leq \sum_{i=1}^n \mathbb{E}\left[\frac{|X_i|^{p-2}}{a^{p-2}}\left(X_i^2 \mathbf{1}\{|X_i| > a\}\right)\right] \leq \sum_{i=1}^n \mathbb{E}|X_i|^p a^{2-p} \leq na^{2-p}K.
\end{aligned}
$$

The desired result follows from the above inequality and equation (44). □

Lemma 4. *Let $\{Y_t\}_{t=1}^T$ be random variables with β-mixing coefficient satisfying $\beta(i) \leq \tau_1 \exp(-\tau_2 i^{\tau_3})$ for some constants $\tau_1, \tau_2, \tau_3 > 0$. Suppose that $\mathbb{E}Y_t = 0$ and $\max_{1\leq t\leq T} \mathbb{E}|Y_t|^p \leq D$ for some constants $p > 2$ and $D > 0$. Then for any $p_0 \in (2, p)$, there exist constants $K_1, ..., K_5 > 0$ such that $\forall w \geq 1$,*

$$
\mathbb{P}\left(\left|\sum_{t=s+1}^{s+n} Y_t\right| \geq \sqrt{n}K_1 w\right) \leq 2\exp\left(-K_2 w^2\right) + K_3 n^{1-p_0/2}\log^{-K_4} n
$$

and

$$
\mathbb{P}\left(\max_{1\leq s\leq T-n}\left|\sum_{t=s+1}^{s+n} Y_t\right| \geq \sqrt{n\log T}K_5\right) \leq 2K_3 T n^{1-p_0/2}\log^{-K_4} n.
$$

Proof. Fix $1 \leq s \leq T - n$. Let $m_1 > m_2$ and $k = \lfloor n/m \rfloor$, where $m = m_1 + m_2$. For $1 \leq j \leq k$, define $H_{j,1} = \{(j-1)m + i : 1 \leq i \leq m_1\}$ and $H_{j,2} = \{(j-1)m + i :\leq m_1 + 1 \leq i \leq m\}$. Also define $H_* = \{km + 1, ..., n\}$. Let $W_{j,1} = m_1^{-1/2}\sum_{t\in H_{j,1}} Y_t$,

$W_{j,2} = m_2^{-1/2} \sum_{t \in H_{j,2}} Y_t$ and $W_* = \sum_{t \in H_*} Y_t$.

Step 1: bound $\sum_{j=1}^k W_{j,1}$.

We apply Lemma 3 together with a Berbee-type coupling result. By Lemma 7.1 of Chen et al. (2016), there exist independent random variables $\{Z_j\}_{j=1}^k$ (possibly on an extended probability space) such that Z_j and $W_{j,1}$ have the same distribution and

$$\mathbb{P}\left(\bigcup_{j=1}^k \{Z_j \neq W_{j,1}\}\right) \leq k\beta(m_2) \leq k\tau_1 \exp\left(-\tau_2 m_2^{\tau_3}\right). \tag{45}$$

Lemma 7.2 of Chen et al. (2016) also implies that there exist constants $M_0, M_1, M_2 > 0$ such that

$$\mathbb{E}|W_{j,1}|^{p_0} \leq M_1 M_2^{p_0} \quad \text{and} \quad \mathbb{E}|W_{j,1}|^2 \leq M_0. \tag{46}$$

Let $Q_{m_1,k,p_0} := \max_{1 \leq j \leq k} \mathbb{E}|W_{j,1}|^{p_0}$. Let $a_T \to \infty$ be a sequence to be chosen later. Applying Lemma 3, we obtain that $\forall t > 0$,

$$\mathbb{P}\left(\left|\sum_{j=1}^k Z_j \mathbf{1}\{|Z_j| \leq a_T\}\right| \geq \sqrt{k}t + k a_T^{1-p_0} Q_{m_1,k,p_0}\right) \leq 2\exp\left[-\frac{t^2}{2\left(a_T^{2-p_0} Q_{m_1,k,p_0} + a_T t k^{-1/2} + k^{-1}\sum_{j=1}^k \mathbb{E}Z_j^2\right)}\right].$$

Hence, by (46), we have

$$\mathbb{P}\left(\left|\sum_{j=1}^k Z_j \mathbf{1}\{|Z_j| \leq a_T\}\right| \geq \sqrt{k}t + k a_T^{1-p_0} M_1 M_2^{p_0}\right) \leq 2\exp\left[-\frac{t^2}{2\left(a_T^{2-p_0} M_1 M_2^{p_0} + a_T t k^{-1/2} + M_0\right)}\right].$$

The above display and (45) imply that

$$\mathbb{P}\left(\left|\sum_{j=1}^k W_{j,1} \mathbf{1}\{|W_{j,1}| \leq a_T\}\right| \geq \sqrt{k}t + k a_T^{1-p_0} M_1 M_2^{p_0}\right)$$

$$\leq 2\exp\left[-\frac{t^2}{2\left(a_T^{2-p_0}M_1M_2^{p_0}+a_Ttk^{-1/2}+M_0\right)}\right]+k\tau_1\exp\left(-\tau_2 m_2^{\tau_3}\right). \quad (47)$$

By (46), we have

$$\mathbb{P}\left(\bigcup_{j=1}^{k}\{|W_{j,1}|\geq a_T\}\right)\leq\sum_{j=1}^{k}\mathbb{P}\left(|W_{j,1}|^{p_0}\geq a_T^{p_0}\right)\leq\sum_{j=1}^{k}\mathbb{E}|W_{j,1}|^{p_0}a_T^{-p_0}\leq ka_T^{-p_0}M_1M_2^{p_0}.$$

By the above two displays,

$$\begin{aligned}&\mathbb{P}\left(\left|\sum_{j=1}^{k}W_{j,1}\right|\geq\sqrt{k}t+ka_T^{1-p_0}M_1M_2^{p_0}\right)\\&\leq 2\exp\left[-\frac{t^2}{2\left(a_T^{2-p_0}M_1M_2^{p_0}+a_Ttk^{-1/2}+M_0\right)}\right]+k\tau_1\exp\left(-\tau_2 m_2^{\tau_3}\right)+ka_T^{-p_0}M_1M_2^{p_0}.\end{aligned} \quad (48)$$

Step 2: bound $\sum_{j=1}^{k}W_{j,2}$ and W_*

Similar to Step 1, we can show that for any $t>0$,

$$\begin{aligned}&\mathbb{P}\left(\left|\sum_{j=1}^{k}W_{j,2}\right|\geq\sqrt{k}t+ka_T^{1-p_0}M_1M_2^{p_0}\right)\\&\leq 2\exp\left[-\frac{t^2}{2\left(a_T^{2-p_0}M_1M_2^{p_0}+a_Ttk^{-1/2}+M_0\right)}\right]+k\tau_1\exp\left(-\tau_2 m_1^{\tau_3}\right)+ka_T^{-p_0}M_1M_2^{p_0}.\end{aligned} \quad (49)$$

Notice that there are fewer than m elements in H_*. Hence, $\forall t>0$,

$$\mathbb{P}\left(|W_*|\geq t\right)\leq\frac{\mathbb{E}|W_*|^{p_0}}{t^{p_1}}\overset{(i)}{\leq}M_1M_2^{p_0}(\sqrt{m}/t)^{p_0}, \quad (50)$$

where (i) follows by Lemma 7.2 of Chen et al. (2016) (with the same constants M_1 and M_2 as in (46)).

Step 3: derive the final result.

Now we choose $m_2=1+\left\lfloor[(p_0/\tau_2)\log n]^{4+1/\tau_3}\right\rfloor$, $m_1=m_2^2$, $m=m_1+m_2$, $k=$

$\lfloor n/m \rfloor$ and $a_T = \sqrt{k}$. Let

$$\begin{aligned} g_n =& k\tau_1 \exp\left(-\tau_2 m_2^{\tau_3}\right) + k a_T^{-p_0} M_1 M_2^{p_0} \\ &+ 2\exp\left[-\frac{n/(m_2 k)}{2\left(a_T^{2-p_0} M_1 M_2^{p_0} + a_T\sqrt{n/(m_2 k)}k^{-1/2} + M_0\right)}\right] \\ &+ k\tau_1 \exp\left(-\tau_2 m_1^{\tau_3}\right) + k a_T^{-p_0} M_1 M_2^{p_0} + M_1 M_2^{p_0}(m/n)^{p_0/2}, \end{aligned}$$

For large n and $z \geq 1$, we have that

$$\begin{aligned} &\mathbb{P}\left(\left|\sum_{t=s+1}^{s+n} Y_t\right| \geq 5\sqrt{n}z\right) \\ =&\mathbb{P}\left(\left|\sqrt{m_1}\sum_{j=1}^{k} W_{j,1} + \sqrt{m_2}\sum_{j=1}^{k} W_{j,2} + W_*\right| \geq 5\sqrt{n}z\right) \\ \leq&\mathbb{P}\left(\left|\sqrt{m_1}\sum_{j=1}^{k} W_{j,1}\right| \geq 2\sqrt{n}z\right) + \mathbb{P}\left(\left|\sqrt{m_2}\sum_{j=1}^{k} W_{j,2}\right| \geq 2\sqrt{n}\right) + \mathbb{P}\left(|W_*| \geq \sqrt{n}\right) \\ \overset{(i)}{\leq}&\mathbb{P}\left(\left|\sum_{j=1}^{k} W_{j,1}\right| \geq \sqrt{k}z + k a_T^{1-p_0} M_1 M_2^{p_0}\right) \\ &+ \mathbb{P}\left(\left|\sum_{j=1}^{k} W_{j,2}\right| \geq \sqrt{k}\sqrt{n/(m_2 k)} + k a_T^{1-p_0} M_1 M_2^{p_0}\right) + \mathbb{P}\left(|W_*| \geq \sqrt{n}\right) \\ \overset{(ii)}{\leq}&\underbrace{2\exp\left[-\frac{z^2}{2\left(a_T^{2-p_0} M_1 M_2^{p_0} + a_T z k^{-1/2} + M_0\right)}\right]}_{\Psi_n(z)} + g_n, \end{aligned}$$

where (i) holds by $z \geq 1 \geq \sqrt{k} a_T^{1-p_0} M_1 M_2^{p_0}$ and $n/m_1 \geq k$ and (ii) follows by (48), (49) and (50). By straight-forward computations, we have that $\Psi_n(z) \leq 2\exp(-M_3 z^2)$ and $g_n \leq n^{1-p_0/2}\log^{-M_4} n$, where M_3 and M_4 are positive constants. This proves the first claim. The second claim follows by the union bound. □

To study properties of estimation errors of the form $(\sum_{s=1}^n X_s \varepsilon_{s+1})/(\sum_{s=1}^n X_s^2)$, we consider the following condition.

Condition 1. Let $\{X_s, e_{s+1}\}_{s=1}^n$ be random variables with β-mixing coefficient satisfying $\beta_{mix}(i) \leq \tau_1 \exp(-\tau_2 i^{\tau_3})$ for some constants $\tau_1, \tau_2, \tau_3 > 0$. Suppose that

$\mathbb{E}X_s e_{s+1} = 0$, $\max_{1\le s\le n}\mathbb{E}|X_s|^p \le D$ and $\max_{1\le s\le n}\mathbb{E}|e_{s+1}|^p \le D$ for some constants $p > 4$ and $D > 0$. Moreover, $D_0 \le \mathbb{E}(n^{-1/2}\sum_{s=1}^n X_s e_{s+1})^2 \le D_1$ and $D_2 \le \mathbb{E}X_s^2 \le D_3$ for some constants $D_0, .., D_3 > 0$.

Lemma 5. *Let Condition 1 hold. Define*

$$\delta = \frac{\sum_{s=1}^n X_s e_{s+1}}{\sum_{s=1}^n X_s^2}.$$

Then for any $p_0 \in (2, p/2)$, there exist $\tilde{\delta} \in \sigma(\{X_s, e_{s+1}\}_{s=1}^n)$ and constants $C_1, ..., C_5 > 0$ such that $\mathbb{P}(\tilde{\delta} \ne \delta) \le C_1 n^{1-p_0/2}\log^{-C_2} n$, $|\mathbb{E}\tilde{\delta}| \le n^{-1}\sqrt{\log n}C_3$ and $n^{-1}C_4 \le \mathbb{E}\tilde{\delta}^2 \le n^{-1}C_5$. Moreover, $|\mathbb{E}\tilde{\delta}^2 - \mathbb{E}(\sum_{s=1}^n X_s e_{s+1})^2/\mathbb{E}(\sum_{s=1}^n X_s^2)^2| \le C_6\sqrt{n^{-3}\log n}$ for some constant $C_6 > 0$.

Proof. Let $Z_{n,1} = n^{-1}\sum_{s=1}^n X_s e_{s+1}$ and $Z_{n,2} = n^{-1}\sum_{s=1}^n X_s^2$. Hence, $\delta = Z_{n,1}/Z_{n,2}$. The proof proceeds in two steps.

Step 1: bound $\mathbb{E}|Z_{n,1}|$.

By Davydov's inequality (Corollary 16.2.4 of Athreya and Lahiri (2006)) and the uniform boundedness of $\mathbb{E}|X_s e_{s+1}|^{2+c}$ for some $c > 0$, we have that for $|s_1 - s_2| > 1$, $|\mathbb{E}X_{s_1}X_{s_2}e_{s_1+1}e_{s_2+1}| \le M_1[\beta(|s_1 - s_2 - 1|)]^{M_2}$ for some constants $M_1, M_2 > 0$. The exponential-decay of the β-mixing coefficient implies that for $|s_1 - s_2| > 1$,

$$|\mathbb{E}X_{s_1}X_{s_2}e_{s_1+1}e_{s_2+1}| \le M_3\exp\left(-M_4|s_1 - s_2 - 1|^{M_5}\right), \tag{51}$$

where $M_3, M_4, M_5 > 0$ are constants. Let $S = \{(s_1, s_2) : 1 \le s_1, s_2 \le n,\ |s_1 - s_2| > 1\}$. Let $M_6 > 0$ be a constant such that $|\mathbb{E}X_{s_1}X_{s_2}e_{s_1+1}e_{s_2+1}| \le M_6$; such a constant exists since $\mathbb{E}|X_s e_{s+1}|^2 \le (\mathbb{E}|X_s|^4\mathbb{E}|e_{s+1}|^4)^{1/2}$ is uniformly bounded. Notice that

$$\begin{aligned}
\mathbb{E}|Z_{n,1}|^2 =& n^{-2}\sum_{s_1=1}^n\sum_{s_2=1}^n \mathbb{E}X_{s_1}X_{s_2}e_{s_1+1}e_{s_2+1} \\
=& n^{-2}\sum_{s=1}^n \mathbb{E}X_s^2 e_{s+1}^2 + n^{-2}\sum_{s=1}^{n-1}\mathbb{E}X_s X_{s+1}e_{s+1}e_{s+2} \\
&+ n^{-2}\sum_{s=2}^{n-1}\mathbb{E}X_s X_{s-1}e_{s+1}e_s + n^{-2}\sum_{(s_1,s_2)\in S}\mathbb{E}X_{s_1}X_{s_2}e_{s_1+1}e_{s_2+1} \\
\le& n^{-1}M_6 + 2n^{-2}(n-1)M_6 + n^{-2}\sum_{(s_1,s_2)\in S}\mathbb{E}X_{s_1}X_{s_2}e_{s_1+1}e_{s_2+1}
\end{aligned}$$

$$
\begin{aligned}
&\overset{(i)}{\leq} n^{-1}M_6 + 2n^{-2}(n-1)M_6 + n^{-2}\sum_{(s_1,s_2)\in S} M_3 \exp\left(-M_4|s_1-s_2-1|^{M_5}\right) \\
&\leq n^{-1}M_6 + 2n^{-2}(n-1)M_6 + n^{-2}\sum_{s_1=1}^{n}\sum_{s_2=1}^{\infty} M_3 \exp\left(-M_4|s_1-s_2-1|^{M_5}\right) \\
&\overset{(ii)}{\leq} n^{-1}M_6 + 2n^{-2}(n-1)M_6 + n^{-2}nM_7 \qquad (52)
\end{aligned}
$$

for some constant $M_7 > 0$, where (i) follows by (51) and (ii) follows by the fact that $\sum_{s_2=1}^{\infty} M_3 \exp(-M_4|s_1 - s_2 - 1|^{M_5})$ is uniformly bounded for any $1 \leq s_1 \leq n$. Hence, there exists a constant $M_8 > 0$ such that

$$
\mathbb{E}|Z_{n,1}| \leq \sqrt{\mathbb{E}|Z_{n,1}|^2} \leq M_8 n^{-1/2}. \qquad (53)
$$

Step 2: derive the desired result.

Notice that $X_s^2 - \mathbb{E}X_s^2$ has uniformly bounded $0.5p$-th moments. By Lemma 4 (applied with $Y_s = X_s^2 - \mathbb{E}X_s^2$), we have that

$$
\mathbb{P}\left(|Z_{n,2} - \mathbb{E}Z_{n,2}| \geq K_1\sqrt{n^{-1}\log n}\right) \leq K_2 n^{1-p_0/2}\log^{-K_3} n,
$$

where $K_1, K_2, K_3 > 0$ are constants. Let $\bar{\delta} = Z_{n,1}/\mathbb{E}Z_{n,2}$ and $\tilde{\delta} = Z_{n,1}/\tilde{Z}_{n,2}$ with

$$
\tilde{Z}_{n,2} = \begin{cases} \mathbb{E}Z_{n,2} + K_1\sqrt{n^{-1}\log n} & \text{if } Z_{n,2} \geq \mathbb{E}Z_{n,2} + K_1\sqrt{n^{-1}\log n} \\ \mathbb{E}Z_{n,2} - K_1\sqrt{n^{-1}\log n} & \text{if } Z_{n,2} \leq \mathbb{E}Z_{n,2} - K_1\sqrt{n^{-1}\log n} \\ Z_{n,2} & \text{otherwise} \end{cases}.
$$

Clearly, $\tilde{\delta} \in \sigma\left(\{X_s, e_{s+1}\}_{s=1}^n\right)$. Moreover,

$$
\mathbb{P}\left(\delta \neq \tilde{\delta}\right) = \mathbb{P}\left(Z_{n,2} \neq \tilde{Z}_{n,2}\right) \leq K_2 n^{1-p_0/2}\log^{-K_3} n. \qquad (54)
$$

Notice that

$$
\begin{aligned}
\mathbb{E}|\tilde{\delta} - \bar{\delta}| = \mathbb{E}\left|\frac{Z_{n,1}}{\tilde{Z}_{n,2}} - \frac{Z_{n,1}}{\mathbb{E}Z_{n,2}}\right| &= \mathbb{E}\left|\frac{Z_{n,1}(\tilde{Z}_{n,2} - \mathbb{E}Z_{n,2})}{\tilde{Z}_{n,2}\mathbb{E}Z_{n,2}}\right| \\
&\overset{(i)}{\leq} \frac{K_1\sqrt{n^{-1}\log n}}{(\mathbb{E}Z_{n,2})\left(\mathbb{E}Z_{n,2} + K_1\sqrt{n^{-1}\log n}\right)}\mathbb{E}\left|Z_{n,1}\right|
\end{aligned}
$$

$$\overset{(ii)}{\leq} K_4 n^{-1}\sqrt{\log n} \qquad \text{for some constant } K_4 > 0,$$

where (i) holds by $|\tilde{Z}_{n,2} - \mathbb{E}Z_{n,2}| \leq K_1\sqrt{n^{-1}\log n}$ (by the definition of $\tilde{Z}_{n,2}$) and (ii) follows by (53) and $\mathbb{E}Z_{n,2} \geq D_2$. Since $\mathbb{E}X_s e_{s+1} = 0$, we have $\mathbb{E}Z_{n,1} = 0$ and $\mathbb{E}\bar{\delta} = 0$. Hence, the above display implies that

$$|\mathbb{E}\tilde{\delta}| \leq |\mathbb{E}\bar{\delta}| + \mathbb{E}|\tilde{\delta} - \bar{\delta}| \leq K_4 n^{-1}\sqrt{\log n}. \tag{55}$$

Lastly, notice that

$$\begin{aligned}\left|\mathbb{E}(\tilde{\delta}^2 - \bar{\delta}^2)\right| &= \mathbb{E}\left[\frac{Z_{n,1}^2|\tilde{Z}_{n,2} - \mathbb{E}Z_{n,2}| \cdot |\tilde{Z}_{n,2} + \mathbb{E}Z_{n,2}|}{\tilde{Z}_{n,2}^2\left(\mathbb{E}Z_{n,2}\right)^2}\right] \\ &\overset{(i)}{\leq} \left\{\frac{K_1\sqrt{n^{-1}\log n}(2\mathbb{E}Z_{n,2} + K_1\sqrt{n^{-1}\log n})}{\left(\mathbb{E}Z_{n,2}\right)^2\left(\mathbb{E}Z_{n,2} - K_1\sqrt{n^{-1}\log n}\right)^2}\right\}\mathbb{E}Z_{n,1}^2 \overset{(ii)}{\leq} K_5\sqrt{n^{-3}\log n} \end{aligned} \tag{56}$$

for some constant $K_5 > 0$, where (i) follows by $|\tilde{Z}_{n,2} - \mathbb{E}Z_{n,2}| \leq K_1\sqrt{n^{-1}\log n}$ (by the definition of $\tilde{Z}_{n,2}$) and (ii) follows by (52) and $\mathbb{E}Z_{n,2} \geq D_2$. Since $\mathbb{E}\bar{\delta}^2 = \mathbb{E}Z_{n,1}^2/(\mathbb{E}Z_{n,2})^2$, $D_2 \leq \mathbb{E}Z_{n,2} \leq D_3$ and $n^{-1}D_0 \leq \mathbb{E}Z_{n,1}^2 \leq n^{-1}D_1$, it follows, by (56), that there exist constants $K_6, K_7 > 0$ such that for large n,

$$n^{-1}K_6 \leq \mathbb{E}\bar{\delta}^2 - \left|\mathbb{E}(\tilde{\delta}^2 - \bar{\delta}^2)\right| \leq \mathbb{E}\tilde{\delta}^2 \leq \mathbb{E}\bar{\delta}^2 + \left|\mathbb{E}(\tilde{\delta}^2 - \bar{\delta}^2)\right| \leq n^{-1}K_7.$$

The desired result follows by (54), (55) and the above display. □

Lemma 6. *Let Condition 1 hold. Define*

$$\delta = \frac{\sum_{s=1}^{n} X_s e_{s+1}}{\sum_{s=1}^{n} X_s^2} \qquad \text{and} \qquad \bar{\delta} = \frac{\sum_{s=1}^{n-a_n} X_s e_{s+1}}{\sum_{s=1}^{n-a_n} X_s^2},$$

where $a_n \leq cn$ for some constant $c \in (0,1)$. Then for any $p_0 \in (2, p/2)$, there exists a constant $M > 0$ such that $\forall x > 0$,

$$\mathbb{P}\left(|\bar{\delta} - \delta| \geq x\right) \leq M \max\left\{n^{1-p_0/2}\log^{-K_3} n,\ (nx/\sqrt{a_n})^{-p_0},\ \left(a_n^{-1}n^{3/2}x/\sqrt{\log n}\right)^{-p/2}\right\}.$$

Proof. After straightforward computations, we have that

$$\delta - \bar{\delta} = \underbrace{\frac{\sum_{s=n-a_n+1}^{n} X_s e_{s+1}}{\sum_{s=1}^{n} X_s^2}}_{J_1} - \underbrace{\frac{\left(\sum_{s=1}^{n-a_n} X_s e_{s+1}\right)\left(\sum_{s=n-a_n+1}^{n} X_s^2\right)}{\left(\sum_{s=1}^{n} X_s^2\right)\left(\sum_{s=1}^{n-a_n} X_s^2\right)}}_{J_2}. \tag{57}$$

Notice that by Lemma 2, both $X_s^2 - \mathbb{E}X_s^2$ and $X_s e_{s+1}$ has uniformly bounded $0.5p$-th moments. Applying Lemma 4 (with $Y_s = X_s^2 - \mathbb{E}X_s^2$ and $Y_s = X_s e_{s+1}$) and using $(1-c)n \leq n - a_n \leq n$, we have that for some constants $K_1, K_2, K_3 > 0$,

$$\begin{cases} \mathbb{P}\left(\left|\sum_{s=1}^{n}(X_s^2 - \mathbb{E}X_s^2)\right| \geq K_1\sqrt{n\log n}\right) \leq K_2 n^{1-p_0/2}\log^{-K_3} n \\ \mathbb{P}\left(\left|\sum_{s=1}^{n-a_n}(X_s^2 - \mathbb{E}X_s^2)\right| \geq K_1\sqrt{n\log n}\right) \leq K_2 n^{1-p_0/2}\log^{-K_3} n \\ \mathbb{P}\left(\left|\sum_{s=1}^{n-a_n}(X_s^2 - \mathbb{E}X_s^2)\right| \geq K_1\sqrt{n\log n}\right) \leq K_2 n^{1-p_0/2}\log^{-K_3} n \\ \mathbb{P}\left(\left|\sum_{s=1}^{n-a_n} X_s e_{s+1}\right| \geq K_1\sqrt{n\log n}\right) \leq K_2 n^{1-p_0/2}\log^{-K_3} n. \end{cases}$$

By Condition 1, $\mathbb{E}X_s^2 \geq D_2$. Since $K_1\sqrt{n^{-1}\log n} < D_2/2$ for large n, we have $\max\left\{\mathbb{P}\left(\sum_{s=1}^{n} X_s^2 \leq nD_2/2\right), \mathbb{P}\left(\sum_{s=1}^{n-a_n} X_s^2 \leq nD_2/2\right)\right\} \leq K_4 n^{1-p_0/2}\log^{-K_3} n$ for some constant $K_4 \geq K_2$. Notice that

$$\begin{aligned} \mathbb{P}(J_1 \geq x/2) &\leq \mathbb{P}\left(\sum_{s=1}^{n} X_s^2 \leq nD_2/2\right) + \mathbb{P}\left(\left|\sum_{s=n-a_n+1}^{n} X_s e_{s+1}\right| \geq D_2 nx/4\right) \\ &\leq \mathbb{P}\left(\sum_{s=1}^{n} X_s^2 \leq nD_2/2\right) + (D_2 nx/4)^{-p_0}\,\mathbb{E}\left|\sum_{s=n-a_n+1}^{n} X_s e_{s+1}\right|^{p_0} \\ &\overset{(i)}{\leq} \mathbb{P}\left(\sum_{s=1}^{n} X_s^2 \leq nD_2/2\right) + (D_2 nx/4)^{-p_0} K_5 a_n^{p_0/2} \text{ for a constant } K_5 > 0 \\ &\leq K_4 n^{1-p_0/2}\log^{-K_3} n + K_5 (D_2 nx/4)^{-p_0} a_n^{p_0/2}, \end{aligned}$$

where (i) holds by Lemma 7.2 of Chen et al. (2016). Also notice that

$$\begin{aligned} &\mathbb{P}(J_2 \geq x/2) \\ &\leq \mathbb{P}\left(\sum_{s=1}^{n-a_n} X_s^2 \leq nD_2/2\right) + \mathbb{P}\left(\sum_{s=1}^{n} X_s^2 \leq nD_2/2\right) \\ &\quad + \mathbb{P}\left(\left|\sum_{s=1}^{n-a_n} X_s e_{s+1}\right| \geq K_1\sqrt{n\log n}\right) + \mathbb{P}\left(\left|\sum_{s=n-a_n+1}^{n} X_s^2\right| \geq \frac{D_2^2 n^2 x}{8K_1\sqrt{n\log n}}\right) \end{aligned}$$

$$
\begin{aligned}
&\leq 3K_4 n^{1-p_0/2}\log^{-K_3} n + \mathbb{P}\left(\left|\sum_{s=n-a_n+1}^{n} X_s^2\right| \geq \frac{D_2^2 n^2 x}{8K_1\sqrt{n\log n}}\right) \\
&\leq 3K_4 n^{1-p_0/2}\log^{-K_3} n + \left(\frac{D_2^2 n^2 x}{8K_1\sqrt{n\log n}}\right)^{-p/2} \mathbb{E}\left|\sum_{s=n-a_n+1}^{n} X_s^2\right|^{p/2} \\
&\overset{(i)}{\leq} 3K_4 n^{1-p_0/2}\log^{-K_3} n + \left(\frac{D_2^2 n^2 x}{8K_1\sqrt{n\log n}}\right)^{-p/2} (a_n D)^{p/2},
\end{aligned}
$$

where (i) follows by $\mathbb{E}|\sum_{s=n-a_n+1}^{n} X_s^2|^{p/2} = \|\sum_{s=n-a_n+1}^{n} X_s^2\|_{L^{p/2}(\mathbb{P})}^{p/2} \leq (\sum_{s=n-a_n+1}^{n} \|X_s^2\|_{L^{p/2}(\mathbb{P})})^{p/2} \leq (a_n D)^{p/2}$. The desired result follows by combining (57) with the above two displays. □

B Proofs of main results

B.1 Proof of Proposition 1

Our proof of Proposition 1 relies on two lemmas, Lemma 7 and 8. We first state and prove these lemmas before proving the proposition.

Lemma 7. *Let Assumption 1 hold. For any constants $K > 0$, $h \in (0,1)$ and $p_1 \in (2, r/2)$, we can enlarge the probability space and construct random variables $\delta_{1,t,*}$, $\delta_{2,t,*}$, $\bar{\delta}_{1,t}$ and $\bar{\delta}_{2,t}$ such that for $j \in \{1,2\}$,*

$$
\begin{cases}
\mathbb{P}\left(\delta_{j,t,*} \neq \delta_{j,t}\right) \leq C_1 T^{-\min\{p_1/2-1,\ (1-h)p_1\}} \\
\mathbb{P}\left(|\delta_{j,t,*} - \bar{\delta}_{j,t}| \leq KT^{-h}\right) = 1 \\
\bar{\delta}_{1,t} \text{ and } \bar{\delta}_{2,t} \text{ are independent of } \{x_{1,s}, x_{2,s}, \varepsilon_{s+1}\}_{s \geq t-1} \\
|\mathbb{E}\bar{\delta}_{j,t}| \leq T^{-1}\sqrt{\log T} C_2 \text{ and } T^{-1}C_3 \leq \mathbb{E}\bar{\delta}_{j,t}^2 \leq T^{-1}C_4 \\
\left|\mathbb{E}\bar{\delta}_{j,t}^2 - \mathbb{E}[\sum_{s=t-n}^{t-1} x_{1,t}(\varepsilon_{t+1} + \beta_2 x_{2,t})]^2 / \mathbb{E}(\sum_{s=t-n}^{t-1} x_{1,t}^2)^2\right| \leq C_5\sqrt{n^{-3}\log n},
\end{cases}
$$

where $C_1, ..., C_5 > 0$ are constants depending only on the constants in Assumption 1.

Proof. We construct $\delta_{j,t,*}$ for $j = 1$; the case for $j = 2$ is analogous. Notice that

$$
\delta_{1,t} = \frac{\sum_{s=t-n}^{t-1} x_{1,t}(\varepsilon_{t+1} + \beta_2 x_{2,t})}{\sum_{s=t-n}^{t-1} x_{1,t}^2}.
$$

Recall the constants in Assumption 1. Let $a_n = \min\{a \in \mathbb{N} \mid a \geq (r/2-1)^{1/c}\log^{1/c} n\}$.

By Theorem 16.2.1 of Athreya and Lahiri (2006), we can extend the probability space with random variables $\{\dot{x}_{1,s}, \dot{x}_{2,s}, \dot{\varepsilon}_{s+1}\}_{s=t-n}^{t-a_n-1}$ such that

$$\begin{cases} \{\dot{x}_{1,s}, \dot{x}_{2,s}, \dot{\varepsilon}_{s+1}\}_{s=t-n}^{t-a_n-1} \text{ has the same distribution as } \{x_{1,s}, x_{2,s}, \varepsilon_{s+1}\}_{s=t-n}^{t-a_n-1} \\ \{\dot{x}_{1,s}, \dot{x}_{2,s}, \dot{\varepsilon}_{s+1}\}_{s=t-n}^{t-a_n-1} \text{ is independent of } \{x_{1,s}, x_{2,s}, \varepsilon_{s+1}\}_{s\geq t-1} \\ \mathbb{P}\left(\{\dot{x}_{1,s}, \dot{x}_{2,s}, \dot{\varepsilon}_{s+1}\}_{s=t-n}^{t-a_n-1} \neq \{x_{1,s}, x_{2,s}, \varepsilon_{s+1}\}_{s=t-n}^{t-a_n-1}\right) = \beta(a_n) \leq b\exp(-a_n^c). \end{cases} \tag{58}$$

Let $\mathcal{F}_n$ be the σ-algebra generated by $\{\dot{x}_{1,s}, \dot{x}_{2,s}, \dot{\varepsilon}_{s+1}\}_{s=t-n}^{t-a_n-1}$. Hence, $\dot{\delta}_{1,t} \in \mathcal{F}_n$ and

$$\mathbb{P}\left(\dot{\delta}_{1,t} \neq \ddot{\delta}_{1,t}\right) \leq b\exp(-a_n^c), \tag{59}$$

where

$$\dot{\delta}_{1,t} = \frac{\sum_{s=t-n}^{t-a_n-1} \dot{x}_{1,t}(\dot{\varepsilon}_{t+1} + \beta_2 \dot{x}_{2,t})}{\sum_{s=t-n}^{t-a_n-1} \dot{x}_{1,t}^2} \text{ and } \ddot{\delta}_{1,t} = \frac{\sum_{s=t-n}^{t-a_n-1} x_{1,t}(\varepsilon_{t+1} + \beta_2 x_{2,t})}{\sum_{s=t-n}^{t-a_n-1} x_{1,t}^2}.$$

Now we apply Lemma 5 (with $X_s = \dot{x}_{1,s}$ and $e_{s+1} = \beta_2 \dot{x}_{2,s} + \dot{\varepsilon}_{s+1}$) and obtain that there exist $\bar{\delta}_{1,t} \in \mathcal{F}_n$ satisfying

$$\begin{cases} \mathbb{P}\left(\dot{\delta}_{1,t} \neq \bar{\delta}_{1,t}\right) \leq M_0 n^{1-p_1/2}\log^{-M_1} n \\ |\mathbb{E}\bar{\delta}_{1,t}| \leq n^{-1}\sqrt{\log n} M_2 \\ n^{-1}M_3 \leq \mathbb{E}\bar{\delta}_{1,t}^2 \leq n^{-1}M_4 \end{cases} \tag{60}$$

where $M_0, ..., M_4 > 0$ are constants. Lemma 5 also implies that $|\mathbb{E}\bar{\delta}_{1,t}^2 - \mathbb{E}[\sum_{s=t-n}^{t-a_n-1} \dot{x}_{1,t}(\dot{\varepsilon}_{t+1} + \beta_2 \dot{x}_{2,t})]^2 / \mathbb{E}(\sum_{s=t-n}^{t-a_n-1} \dot{x}_{1,t}^2)^2| \leq G\sqrt{(n-a_n)^{-3}\log(n-a_n)}$ for some constant $G > 0$. Notice that $a_n \asymp \log^{1/c} n$. In computing this expectation, we can replace $\{\dot{x}_{1,s}, \dot{x}_{2,s}, \dot{\varepsilon}_{s+1}\}_{s=t-n}^{t-a_n-1}$ with $\{x_{1,s}, x_{2,s}, \varepsilon_{s+1}\}_{s=t-n}^{t-a_n-1}$ since they have the same distribution. It is not hard to verify that $|\mathbb{E}\bar{\delta}_{1,t}^2 - \mathbb{E}[\sum_{s=t-n}^{t-1} x_{1,t}(\varepsilon_{t+1} + \beta_2 x_{2,t})]^2 / \mathbb{E}(\sum_{s=t-n}^{t-1} x_{1,t}^2)^2| \leq G\sqrt{n^{-3}\log n}$ for some constant $G' > 0$.

Since $\bar{\delta}_{1,t} \in \mathcal{F}_n$, (58) implies that

$$\bar{\delta}_{1,t} \text{ is independent of } \{x_{1,s}, x_{2,s}, \varepsilon_{s+1}\}_{s\geq t-1}. \tag{61}$$

By Lemma 6 (applied with $X_s = x_{1,s}$ and $e_{s+1} = \beta_2 x_{2,s} + \varepsilon_{s+1}$), we have that for $x = Kn^{-h}$,

$$\mathbb{P}\left(|\ddot{\delta}_{1,t} - \delta_{1,t}| \geq x\right) \leq M_5 \max\left\{n^{1-p_1/2}\log^{-K_3} n,\ (nx/\sqrt{a_n})^{-p_1},\ \left(a_n^{-1}n^{3/2}x/\sqrt{\log n}\right)^{-r/2}\right\}, \quad (62)$$

where $M_5 > 0$ is a constant. Define

$$\delta_{1,t,*} = \begin{cases} \bar{\delta}_{1,t} + x & \text{if } \delta_{1,t} \geq \bar{\delta}_{1,t} + x \\ \bar{\delta}_{1,t} - x & \text{if } \delta_{1,t} \leq \bar{\delta}_{1,t} - x \\ \delta_{1,t} & \text{otherwise} \end{cases}.$$

Hence,

$$|\delta_{1,t,*} - \bar{\delta}_{1,t}| \leq x. \quad (63)$$

Notice that, for some constant $M_6 > 0$, we have

$$\begin{aligned}
&\mathbb{P}\left(\delta_{1,t,*} \neq \delta_{1,t}\right) \\
&\leq \mathbb{P}\left(|\delta_{1,t} - \ddot{\delta}_{1,t}| \geq x\right) + \mathbb{P}\left(\ddot{\delta}_{1,t} \neq \dot{\delta}_{1,t}\right) + \mathbb{P}\left(\dot{\delta}_{1,t} \neq \bar{\delta}_{1,t}\right) \\
&\overset{(i)}{\leq} M_6 \max\left\{n^{1-p_1/2}\log^{-K_3} n,\ (nx/\sqrt{a_n})^{-p_1},\ \left(a_n^{-1}n^{3/2}x/\sqrt{\log n}\right)^{-r/2},\ \exp(-a_n^c)\right\} \\
&\overset{(ii)}{\leq} M_6 \max\left\{n^{1-p_1/2},\ (nx)^{-p_1},\ \left(n^{3/2}x\right)^{-r/2}\right\} \\
&\overset{(iii)}{\leq} M_6 \max\left\{n^{1-p_1/2},\ (nx)^{-p_1}\right\} \leq M_7 n^{-\min\{p_1/2-1,\ (1-h)p_1\}} \text{ for some constant } M_7 > 0,
\end{aligned}$$

where (i) holds by (59), (60) and (62), (ii) follows by the fact that $\exp(-a_n^c) \leq n^{1-r/2} < n^{1-p_1/2}$ and (iii) follows by $(n^{3/2}x)^{-r/2} < (n^{3/2}x)^{-p_1} < (nx)^{-p_1}$ (due to $p_1 < r/2$).

Since $n \asymp T$, the desired result follows by the above display, (63), (61) and (60). □

Lemma 8. *Let Assumption 1 hold. Then $\forall p_1 \in (2, r/2)$ and $\forall h \in (2\alpha_{x,2}, 1)$, there exist constants $G_1, ..., G_4 > 0$ and an array of random variables $\{\Delta L_{t+1,*}\}_{t=n}^{T-n}$ such*

that for $T \geq G_1$,

$$\mathbb{P}\left(\bigcap_{t=n}^{T-1}\{\Delta L_{t+1,*} = \Delta L_{t+1}\}\right) \geq 1 - G_2 T^{1-\min\{p_1/2-1,\ (1-h)p_1\}}$$

and

$$G_3 T^{-2\alpha_{x,2}} \leq \mathbb{E}\Delta L_{t+1,*} \leq G_4 T^{-2\alpha_{x,2}}.$$

Proof. Let $\delta_{j,t} = \hat{\beta}_{j,t} - \beta_{j,t}$. Recall from (11) that

$$\begin{aligned}\Delta L_{t+1} &= 2\varepsilon_{t+1}\left(\beta_2 x_{2,t} - \beta_1 x_{1,t} - \delta_{1,t}x_{1,t} + \delta_{2,t}x_{2,t}\right) \\ &\quad + \left(\beta_2 x_{2,t} - \beta_1 x_{1,t} - \delta_{1,t}x_{1,t} + \delta_{2,t}x_{2,t}\right)\left(\beta_1 x_{1,t} + \beta_2 x_{2,t} - \delta_{1,t}x_{1,t} - \delta_{2,t}x_{2,t}\right). \quad (64)\end{aligned}$$

Let $\delta_{j,t,*}$ and $\bar{\delta}_{j,t}$ be defined as in the statement of Lemma 7 (with $K = 1$):

$$\begin{cases} \mathbb{P}\left(\delta_{j,t,*} \neq \delta_{j,t}\right) \leq C_1 T^{-\min\{p_1/2-1,\ (1-h)p_1\}} \\ \mathbb{P}\left(|\delta_{j,t,*} - \bar{\delta}_{j,t}| \leq T^{-h}\right) = 1 \\ \bar{\delta}_{1,t} \text{ and } \bar{\delta}_{2,t} \text{ are independent of } \{x_{1,s}, x_{2,s}, \varepsilon_{s+1}\}_{s \geq t-1} \\ |\mathbb{E}\bar{\delta}_{j,t}| \leq T^{-1}\sqrt{\log T}C_2 \text{ and } T^{-1}C_3 \leq \mathbb{E}\bar{\delta}_{j,t}^2 \leq T^{-1}C_4, \end{cases} \quad (65)$$

where $C_1, C_2, C_3, C_4 > 0$ are constants. Define

$$\begin{aligned}\Delta L_{t+1,*} &= 2\varepsilon_{t+1}\left(\beta_2 x_{2,t} - \beta_1 x_{1,t} - \delta_{1,t,*}x_{1,t} + \delta_{2,t,*}x_{2,t}\right) \\ &+ \left(\beta_2 x_{2,t} - \beta_1 x_{1,t} - \delta_{1,t,*}x_{1,t} + \delta_{2,t,*}x_{2,t}\right)\left(\beta_1 x_{1,t} + \beta_2 x_{2,t} - \delta_{1,t,*}x_{1,t} - \delta_{2,t,*}x_{2,t}\right). \quad (66)\end{aligned}$$

The first statement in (65) implies that

$$\begin{aligned}\mathbb{P}\left(\bigcap_{t=n}^{T-1}\{\Delta L_{t+1,*} = \Delta L_{t+1}\}\right) &\geq 1 - \sum_{t=n}^{T-1}\mathbb{P}\left(\delta_{1,t} \neq \delta_{1,t,*} \text{ or } \delta_{2,t} \neq \delta_{2,t,*}\right) \\ &\geq 1 - 2C_1(T-n)T^{-\min\{p_1/2-1,\ (1-h)p_1\}}.\end{aligned}$$

Since $T - n < T$, the first claim follows.

Now we compute $\mathbb{E}\Delta L_{t+1,*}$. Notice that there exist a constant $K_1 > 0$ such that for $j_1, j_2 \in \{1, 2\}$,

$$|\mathbb{E}\delta_{j_1,t,*}x_{j_1,t}\varepsilon_{t+1}| \leq |\mathbb{E}\bar{\delta}_{j_1,t}x_{j_1,t}\varepsilon_{t+1}| + \mathbb{E}|(\delta_{j_1,t,*} - \bar{\delta}_{j_1,t})x_{j_1,t}\varepsilon_{t+1}|$$

$$\overset{(i)}{=} \mathbb{E}|(\delta_{j_1,t,*} - \bar{\delta}_{j_1,t})x_{j_1,t}\varepsilon_{t+1}|,$$
$$\overset{(ii)}{\leq} T^{-h}\mathbb{E}|x_{j_1,t}\varepsilon_{t+1}| \leq K_1 T^{-h},$$

where (i) follows by the fact that $\bar{\delta}_{j_1,t}$ is independent of $x_{j_1,t}\varepsilon_{t+1}$ and $\mathbb{E}x_{j_1,t}\varepsilon_{t+1} = 0$ and (ii) follows by second statement in (65). Similarly, we have that for some constants $K_2, K_3, K_4 > 0$,

$$\begin{aligned}|\mathbb{E}\delta_{j_1,t,*}x_{j_1,t}x_{j_2,t}| &\leq |\mathbb{E}\bar{\delta}_{j_1,t}x_{j_1,t}x_{j_2,t}| + \mathbb{E}|(\delta_{j_1,t,*} - \bar{\delta}_{j_1,t})x_{j_1,t}x_{j_2,t}| \\ &\overset{(i)}{=} |\mathbb{E}\bar{\delta}_{j_1,t}| \cdot |\mathbb{E}x_{j_1,t}x_{j_2,t}| + \mathbb{E}|(\delta_{j_1,t,*} - \bar{\delta}_{j_1,t})x_{j_1,t}x_{j_2,t}| \\ &\overset{(ii)}{\leq} T^{-1}\sqrt{\log T}C_2K_2 + T^{-h}K_3 \overset{(iii)}{\leq} T^{-h}K_4,\end{aligned}$$

where (i) follows by the independence between $\bar{\delta}_{j_1,t}$ and $\{x_{j_1}, x_{j_2}\}$, (ii) follows by the second and third statements in (65) and (iii) follows by $h < 1$. Moreover, we have that for constants $K_5 \geq \mathbb{E}|x_{j_1,t}x_{j_2,t}|$ and K_6 large enough,

$$\begin{aligned}&|\mathbb{E}\delta_{j_1,t,*}\delta_{j_2,t,*}x_{j_1,t}x_{j_2,t}| \\ &\leq |\mathbb{E}\bar{\delta}_{j_1,t}\bar{\delta}_{j_2,t}x_{j_1,t}x_{j_2,t}| + \mathbb{E}|(\delta_{j_1,t,*} - \bar{\delta}_{j_1,t})\delta_{j_2,t,*}x_{j_1,t}x_{j_2,t}| + \mathbb{E}|(\delta_{j_2,t,*} - \bar{\delta}_{j_2,t})\bar{\delta}_{j_1,t}x_{j_1,t}x_{j_2,t}| \\ &\overset{(i)}{\leq} |\mathbb{E}\bar{\delta}_{j_1,t}\bar{\delta}_{j_2,t}| \cdot \mathbb{E}|x_{j_1,t}x_{j_2,t}| + T^{-h}\mathbb{E}|\delta_{j_2,t,*}x_{j_1,t}x_{j_2,t}| + T^{-h}\mathbb{E}|\delta_{j_1,t,*}x_{j_1,t}x_{j_2,t}| \\ &\overset{(ii)}{\leq} |\mathbb{E}\bar{\delta}_{j_1,t}\bar{\delta}_{j_2,t}| \cdot \mathbb{E}|x_{j_1,t}x_{j_2,t}| + T^{-h}\mathbb{E}\left[(|\bar{\delta}_{j_2,t}| + T^{-h})|x_{j_1,t}x_{j_2,t}|\right] \\ &\quad + T^{-h}\mathbb{E}\left[(|\bar{\delta}_{j_1,t}| + T^{-h})|x_{j_1,t}x_{j_2,t}|\right] \\ &\overset{(iii)}{\leq} |\mathbb{E}\bar{\delta}_{j_1,t}\bar{\delta}_{j_2,t}| \cdot \mathbb{E}|x_{j_1,t}x_{j_2,t}| + T^{-h}\mathbb{E}(|\bar{\delta}_{j_2,t}| + T^{-h}) \cdot \mathbb{E}|x_{j_1,t}x_{j_2,t}| \\ &\quad + T^{-h}\mathbb{E}(|\bar{\delta}_{j_1,t}| + T^{-h}) \cdot \mathbb{E}|x_{j_1,t}x_{j_2,t}| \\ &\overset{(iv)}{\leq} T^{-1}C_4K_5 + 2T^{-h}(T^{-1/2}C_4^{1/2} + T^{-h})K_5 \\ &\leq K_6 \max\left\{T^{-h-1/2},\ T^{-2h}\right\},\end{aligned}$$

where (i), (ii) and (iii) follow by computations based on the independence between $\{\bar{\delta}_{j_1,t}, \bar{\delta}_{j_2,t}\}$ and $\{x_{j_1,t}, x_{j_2,t}\}$ and $|\delta_{j,t,*} - \bar{\delta}_{j,t}| \leq T^{-h}$, while (iv) follows by (65), i.e., $|\mathbb{E}\bar{\delta}_{j_1,t}\bar{\delta}_{j_2,t}| \leq [(\mathbb{E}\bar{\delta}_{j_1,t}^2)(\mathbb{E}\bar{\delta}_{j_2,t}^2)]^{1/2} \leq T^{-1}C_4$ and $\mathbb{E}|\bar{\delta}_{j,t}| \leq (\mathbb{E}\bar{\delta}_{j,t}^2)^{1/2} \leq T^{-1/2}C_4^{1/2}$.

By straight-forward computations based on the previous three displays and (66),

we have that for some constant $K_7 > 0$,

$$\mathbb{E}\left|\Delta L_{t+1,*} - \left[2\varepsilon_{t+1}\left(\beta_2 x_{2,t} - \beta_1 x_{1,t}\right) + \beta_2^2 x_{2,t}^2 - \beta_{1,t}^2 x_{1,t}^2\right]\right| \leq K_7 T^{-h}. \tag{67}$$

Since $\mathbb{E}\varepsilon_{t+1}\beta_j x_{j,t} = 0$, we have that for some constant $K_8 > 0$,

$$\left|\mathbb{E}\Delta L_{t+1,*} - \left(\beta_1^2 \mathbb{E}x_{1,t}^2 - \beta_2^2 \mathbb{E}x_{2,t}^2\right)\right| \leq K_8 T^{-h}.$$

Since $\alpha_{x,2} < \alpha_{x,1}$ (by Assumption 1), $\beta_2^2 \mathbb{E}x_{2,t}^2 - \beta_{1,t}^2 \mathbb{E}x_{1,t}^2 \geq K_9 T^{-2\alpha_{x,2}}$ for some constant $K_9 > 0$. Since $h > 2\alpha_{x,2}$, we have that for large T,

$$K_{10} T^{-2\alpha_{x,2}} \leq \mathbb{E}\Delta L_{t+1,*} \leq K_{11} T^{-2\alpha_{x,2}},$$

where $K_{10}, K_{11} > 0$ are constants. This proves the second claim. □

Proof of Proposition 1. Let $p_1 = 2 + r/4$ and $h = \alpha_{x,2} + 1/2$. Since $r > 8$ and $\alpha_{x,2} \in [0, 1/2)$, we have that $p_1 \in (2, r/2)$ and $h \in (2\alpha_{x,2}, 1)$. Applying Lemma 8, we obtain that there exist constants $G_1, ..., G_4 > 0$ and an array of random variables $\{\Delta L_{t+1,*}\}_{t=n}^{T-n}$ such that for $T \geq G_1$,

$$\mathbb{P}\left(\bigcap_{t=n}^{T-1} \{\Delta L_{t+1,*} = \Delta L_{t+1}\}\right) \geq 1 - G_2 T^{1-\min\{p_1/2-1,\ (1-h)p_1\}}$$

and

$$G_3 T^{-2\alpha_{x,2}} \leq \mathbb{E}\Delta L_{t+1,*} \leq G_4 T^{-2\alpha_{x,2}}.$$

Notice that

$$\begin{aligned}
1 - \min\{p_1/2 - 1,\ (1-h)p_1\} &= \max\{2 - p_1/2,\ 1 + (h-1)p_1\} \\
&= \max\{2 - p_1/2,\ 1 + (\alpha_{x,2} - 1/2)p_1\} \\
&= \max\{1 - r/8,\ 1 + (\alpha_{x,2} - 1/2)(2 + r/4)\}
\end{aligned}$$

The proof is complete. □

B.2 Proof of Proposition 2

Our proof of Proposition 2 relies on two lemmas, Lemma 9 and 10. We first state and prove these lemmas before proving Proposition 2.

Lemma 9. *Suppose that Assumptions 1 and 2 hold. Let $\Delta L_{t+1,*}$ be defined as in (66) in the proof of Lemma 8. Let $\tilde{\theta}_t = (\tilde{\theta}_{1,t}, \tilde{\theta}_{2,t})' = (\sum_{s=t-m}^{t-1} z_s z_s')^{-1}(\sum_{s=t-m}^{t-1} z_s \Delta L_{t+1,*})$, where $z_s = (1, z_{1s})'$. Fix $p \in (2, r/3)$. Then there exist constants $G_0, G_1, G_2, G_3 > 0$ such that for $T \geq G_0$,*

$$\mathbb{P}\left(\left|\frac{\tilde{\theta}_{1,t}}{\tilde{\theta}_{2,t}}\right| \geq G_1 T^{\alpha_{z,2}-\alpha_{x,2}}\right) \leq G_2 T^{1-p/2}\log^{-G_3}$$

and

$$\mathbb{P}\left(\tilde{\theta}_{2,t} \leq 0\right) \leq G_2 T^{1-p/2}\log^{-G_3}.$$

Proof. Let $\Psi_{t+1} = m^{\alpha_{x,2}}\Delta L_{t+1,*}$, $\pi_t = (\sum_{s=t-m}^{t-1}\mathbb{E}z_s z_s')^{-1}(\sum_{s=t-m}^{t-1}\mathbb{E}z_s\Psi_{s+1})$, $\hat{\pi}_t = (\sum_{s=t-m}^{t-1} z_s z_s')^{-1}(\sum_{s=t-m}^{t-1} z_s\Psi_{s+1})$ and $\{\xi_{s+1}\}_{s=t-m}^{t-1}$ with $\xi_{s+1} = \Psi_{s+1} - z_s'\pi_t$. Clearly, $\hat{\pi}_t = m^{\alpha_{x,2}}\tilde{\theta}_t$. Let $\gamma_t = \hat{\pi}_t - \pi_t$. The proof proceeds in two steps. We first bound γ_t and then show the desired results.

Step 1: bound γ_t

By simple computation, we have

$$\gamma_t = \underbrace{\left[m^{-1}\sum_{s=t-m}^{t-1} z_s z_s'\right]^{-1}}_{J_t} \cdot \underbrace{m^{-1}\sum_{s=t-m}^{t-1} z_s\xi_{s+1}}_{B_t}.$$

Since entries of $z_s z_s' - \mathbb{E}z_s z_s'$ has uniformly bounded $0.5r$-th moments, it follows, by Lemma 4, that for some constants $K_1, K_2, K_3 > 0$,

$$\mathbb{P}\left(\left\|m^{-1}\sum_{s=t-m}^{t-1}(z_s z_s' - \mathbb{E}z_s z_s')\right\|_\infty \geq K_1\sqrt{m^{-1}\log m}\right) \leq K_2 m^{1-p/2}\log^{-K_3} m.$$

Since $m^{-1}\sum_{s=t-m}^{t-1}\mathbb{E}z_s z_s' = \mathrm{diag}(1, m^{-1}\sum_{s=t-m}^{t-1}\mathbb{E}z_{1s}^2)$ and $\mathbb{E}z_{1s}^2$ is bounded away from zero and infinity, the eigenvalues of $m^{-1}\sum_{s=t-m}^{t-1}\mathbb{E}z_s z_s'$ lie in $[K_4, K_5]$ for some constants $K_4, K_5 > 0$. By the above display, the eigenvalues of $m^{-1}\sum_{s=t-m}^{t-1} z_s z_s'$ lie in $[K_4/2, 2K_5]$ for large n with probability at least $1 - K_2 m^{1-p/2}\log^{-K_3} m$. Hence, for

some constants $K_6, K_7 > 0$, we have that for $n \geq K_6$,

$$\mathbb{P}\left(\|J_t\|_\infty \geq K_7\right) \leq K_2 m^{1-p/2} \log^{-K_3} m.$$

Recall the definition of $\Delta L_{t+1,*}$ (in (66) in the proof of Lemma 8). We apply Lemma 2. By straight-forward computations, we have that Ψ_{t+1} has uniformly bounded $0.5r$-th moments. Hence, $\xi_{s+1} = \Psi_{s+1} - z_s'\pi_s$ has uniformly bounded $0.5r$-th moments. Again by Lemma 2, entries of $z_s\xi_{s+1}$ has uniformly bounded $\frac{r}{3}$-th moments. Notice that $p \in (2, r/3)$. It follows by Lemma 4 (applied to each entry of $z_s\xi_{s+1}$) that for some constants $K_8, K_9, K_{10} > 0$, we have

$$\mathbb{P}\left(\|B_t\|_\infty \geq K_8\sqrt{m^{-1}\log m}\right) \leq K_9 m^{1-p/2} \log^{-K_{10}} m.$$

It follows by the above two displays that for large m,

$$\begin{aligned}
\mathbb{P}\left(\|\gamma_t\|_\infty \geq 2K_7K_8\sqrt{m^{-1}\log m}\right) &\leq \mathbb{P}\left(2\|J_t\|_\infty\|B_t\|_\infty \geq 2K_7K_8\sqrt{m^{-1}\log m}\right) \\
&\leq \mathbb{P}\left(\|J_t\|_\infty \geq K_7\right) + \mathbb{P}\left(\|B_t\|_\infty \geq K_8\sqrt{m^{-1}\log m}\right) \\
&\leq K_2 m^{1-p/2}\log^{-K_3} m + K_9 m^{1-p/2}\log^{-K_{10}} m.
\end{aligned}$$

Since $T \asymp m$, there are constants $K_{11}, ..., K_{14} > 0$ such that for $T \geq K_{11}$,

$$\mathbb{P}\left(\|\gamma_t\|_\infty \geq K_{12}\sqrt{T^{-1}\log T}\right) \leq K_{13} T^{1-p/2}\log^{-K_{14}}. \tag{68}$$

Step 2: show the desired results.

Partition $\pi_t = (\pi_{1,t}, \pi_{2,t})'$. By $\mathbb{E}z_{1s} = 0$, it follows that $\pi_{1,t} = m^{-1}\sum_{s=t-m}^{t-1}\mathbb{E}m^{\alpha_{x,2}}\Delta L_{s+1,*}$ and $\pi_{2,t} = (\sum_{s=t-m}^{t-1}\mathbb{E}m^{\alpha_{x,2}}z_{1s}\Delta L_{s+1,*})/(m^{-1}\sum_{s=t-m}^{t-1}\mathbb{E}z_{1s}^2)$. By Lemma 8, there are constants $K_{15}, K_{16} > 0$ such that

$$K_{15}T^{-\alpha_{x,2}} \leq \pi_{1,t} \leq K_{16}T^{-\alpha_{x,2}}. \tag{69}$$

By computations similar to (67) in the proof of Lemma 8, one can show that

$$\mathbb{E}\left|z_{1t}\Delta L_{t+1,*} - z_{1t}\left[2\varepsilon_{t+1}\left(\beta_2 x_{2,t} - \beta_1 x_{1,t}\right) + \beta_2^2 x_{2,t}^2 - \beta_{1,t}^2 x_{1,t}^2\right]\right| \leq K_{17}T^{-1/2}\sqrt{\log T},$$

where $K_{17} > 0$ is a constant. By Assumption 1,

$$\mathbb{E}\left\{z_{1t}\left[2\varepsilon_{t+1}\left(\beta_2 x_{2,t} - \beta_1 x_{1,t}\right) + \beta_2^2 x_{2,t}^2 - \beta_{1,t}^2 x_{1,t}^2\right]\right\} \geq K_{18} T^{-\alpha_{x,2}-\alpha_{z,2}}.$$

It follows that for large T, $\mathbb{E}z_{1t}\Delta L_{t+1,*} \geq K_{18}T^{-\alpha_{x,2}-\alpha_{z,2}}/2$. Hence, for some constant $K_{19} > 0$,

$$\pi_{2,t} \geq K_{19} T^{-\alpha_{z,2}}. \tag{70}$$

Let $x = 2K_{16}T^{-\alpha_{x,2}}$ and $M = 2x/(K_{19}T^{-\alpha_{z,2}})$. Then

$$\begin{aligned}
&\mathbb{P}\left(|\hat{\pi}_{1,t}| \geq M|\hat{\pi}_{2,t}|\right) \\
&\leq \mathbb{P}\left(|\hat{\pi}_{1,t}| \geq x\right) + \mathbb{P}\left(|\hat{\pi}_{2,t}| \leq x/M\right) \\
&\leq \mathbb{P}\left(|\hat{\pi}_{1,t} - \pi_{1,t}| \geq x - |\pi_1|\right) + \mathbb{P}\left(|\hat{\pi}_{2,t} - \pi_{2,t}| \geq |\pi_{2,t}| - x/M\right) \\
&\overset{(i)}{\leq} \mathbb{P}\left(|\hat{\pi}_{1,t} - \pi_{1,t}| \geq x - K_{16}T^{-\alpha_{x,2}}\right) + \mathbb{P}\left(|\hat{\pi}_{2,t} - \pi_{2,t}| \geq K_{19}T^{-\alpha_{z,2}} - x/M\right) \\
&\leq \mathbb{P}\left(\|\gamma_t\|_\infty \geq x - K_{16}T^{-\alpha_{x,2}}\right) + \mathbb{P}\left(\|\gamma_t\|_\infty \geq K_{19}T^{-\alpha_{z,2}} - x/M\right) \\
&= \mathbb{P}\left(\|\gamma_t\|_\infty \geq K_{16}T^{-\alpha_{x,2}}\right) + \mathbb{P}\left(\|\gamma_t\|_\infty \geq K_{19}T^{-\alpha_{z,2}}/2\right) \\
&\overset{(ii)}{\leq} 2K_{13}T^{1-p/2}\log^{-K_{14}},
\end{aligned}$$

where (i) holds by (69) and (70) and (ii) follows by (68) together with $T^{-\alpha_{x,2}} \gg \sqrt{T^{-1}\log T}$ and $T^{-\alpha_{z,2}} \gg \sqrt{T^{-1}\log T}$. The first claim follows by $|\hat{\pi}_{1,t}/\hat{\pi}_{2,t}| = |\tilde{\theta}_{1,t}/\tilde{\theta}_{2,t}|$ and $M = 2x/(K_{19}T^{-\alpha_{z,2}}) = (4K_{16}/K_{19})T^{\alpha_{z,2}-\alpha_{x,2}}$.

To see the second claim, notice that

$$\begin{aligned}
\mathbb{P}\left(\tilde{\theta}_{2,t} \leq 0\right) \overset{(i)}{=} \mathbb{P}\left(\hat{\pi}_{2,t} \leq 0\right) = \mathbb{P}\left(\gamma_{2,t} \leq -\pi_{2,t}\right) &\overset{(ii)}{\leq} \mathbb{P}\left(\gamma_{2,t} \leq -K_{19}T^{-\alpha_{z,2}}\right) \\
&\leq \mathbb{P}\left(\|\gamma_t\|_\infty \geq K_{19}T^{-\alpha_{z,2}}\right) \\
&\overset{(iii)}{\leq} K_{13}T^{1-p/2}\log^{-K_{14}},
\end{aligned}$$

where (i) holds by $\tilde{\theta}_t = m^{-\alpha_{x,2}}\hat{\pi}_t$, (ii) follows by (70) and (iii) holds by (68) and $T^{-\alpha_{z,2}} \gg \sqrt{T^{-1}\log T}$. □

Lemma 10. *Let Assumptions 1 and 2 hold. Fix any $p_1 \in (2, r/2)$ and $h \in (2\alpha_{x,2}, 1)$. Then there exist constants $G_0, G_1, ..., G_5 > 0$ and an array $\{S_{t+1}\}_{t=n+m}^{T-1}$ such that*

$\forall T \geq G_0$,

$$\mathbb{P}\left(\bigcap_{t=n+m}^{T-1}\left\{S_{t+1} = \Delta L_{t+1}\mathbf{1}\{z_t'\hat{\theta}_{m,t} > 0\}\right\}\right) \geq 1 - G_2 T^{1-\min\{p_1/2-1,\ (1-h)p_1\}}$$

and

$$\mathbb{E}S_{t+1} \geq K_5 T^{-\alpha_{x,2}-\alpha_{z,2}}.$$

Proof. Let $\tilde{\theta}_t = (\sum_{s=t-m}^{t-1} z_s z_s')^{-1}(\sum_{s=t-m}^{t-1} z_s \Delta L_{t+1,*})$, where $\Delta L_{t+1,*}$ is defined as in (66) in the proof of Lemma 8. Define $S_{t+1} = \Delta L_{t+1,*}\mathbf{1}\{z_t'\tilde{\theta}_t > 0\}$. Notice that

$$\bigcap_{t=n}^{T-1}\{\Delta L_{t+1,*} = \Delta L_{t+1}\} \subseteq \bigcap_{t=n+m}^{T-1}\left\{S_{t+1} = \Delta L_{t+1}\mathbf{1}\{z_t'\hat{\theta}_{m,t} > 0\}\right\}.$$

Hence, the first claim follows by Lemma 8.

It remains to bound $\mathbb{E}S_{t+1}$. To this end, let $q = r/2$ and $\nu = r/(r-2)$. Hence, $q, \nu > 1$ and $q^{-1} + \nu^{-1} = 1$. Notice that

$$\begin{aligned}
&\mathbb{E}\left(|\Delta L_{t+1,*}|\left|\mathbf{1}\{z_t'\tilde{\theta}_t > 0\} - \mathbf{1}\{z_{1t} > 0\}\right|\right) \\
&= \mathbb{E}\Bigg(|\Delta L_{t+1,*}| \cdot \Big|\mathbf{1}\{z_{1t} > -\tilde{\theta}_{1,t}/\tilde{\theta}_{2,t} \text{ and } \tilde{\theta}_{2,t} > 0\} \\
&\qquad\qquad + \mathbf{1}\{z_{1t} < -\tilde{\theta}_{1,t}/\tilde{\theta}_{2,t} \text{ and } \tilde{\theta}_{2,t} \leq 0\} - \mathbf{1}\{z_{1t} > 0\}\Big|\Bigg) \\
&\leq \mathbb{E}\left(|\Delta L_{t+1,*}|\left[\mathbf{1}\left\{0 < z_{1t} \leq -\tilde{\theta}_{1,t}/\tilde{\theta}_{2,t}\right\} + \mathbf{1}\left\{\tilde{\theta}_{2,t} \leq 0\right\}\right]\right) \\
&\overset{(i)}{\leq} \|\Delta L_{t+1,*}\|_{L^q(\mathbb{P})}\left\|\mathbf{1}\left\{0 < z_{1t} \leq -\tilde{\theta}_{1,t}/\tilde{\theta}_{2,t}\right\} + \mathbf{1}\left\{\tilde{\theta}_{2,t} \leq 0\right\}\right\|_{L^\nu(\mathbb{P})} \\
&\leq \|\Delta L_{t+1,*}\|_{L^q(\mathbb{P})}\left[\left\|\mathbf{1}\left\{0 < z_{1t} \leq -\tilde{\theta}_{1,t}/\tilde{\theta}_{2,t}\right\}\right\|_{L^\nu(\mathbb{P})} + \left\|\mathbf{1}\left\{\tilde{\theta}_{2,t} \leq 0\right\}\right\|_{L^\nu(\mathbb{P})}\right] \\
&= \|\Delta L_{t+1,*}\|_{L^q(\mathbb{P})}\left(\left[\mathbb{P}\left(0 < z_{1t} \leq -\tilde{\theta}_{1,t}/\tilde{\theta}_{2,t}\right)\right]^{1/\nu} + \left[\mathbb{P}\left(\tilde{\theta}_{2,t} \leq 0\right)\right]^{1/\nu}\right) \qquad (71)
\end{aligned}$$

where (i) follows by Holder's inequality. By Assumption 2, the p.d.f of z_{1t} in a fixed neighborhood of zero is bounded above by some constant $K_0 > 0$. Recall constants $G_1, ..., G_4 > 0$ in the statement of Lemma 9. Hence,

$$\mathbb{P}\left(0 < z_{1t} \leq -\tilde{\theta}_{1,t}/\tilde{\theta}_{2,t}\right) \leq \mathbb{P}\left(0 < z_{1t} \leq \left|\tilde{\theta}_{1,t}/\tilde{\theta}_{2,t}\right|\right)$$

$$\leq \mathbb{P}\left(0 < z_{1t} \leq G_1 T^{\alpha_{z,2}-\alpha_{x,2}}\right) + \mathbb{P}\left(\left|\tilde{\theta}_{1,t}/\tilde{\theta}_{2,t}\right| \geq G_1 T^{\alpha_{z,2}-\alpha_{x,2}}\right)$$
$$\overset{(i)}{\leq} K_0 G_1 T^{\alpha_{z,2}-\alpha_{x,2}} + G_2 T^{1-p/2} \log^{-G_3}, \tag{72}$$

where (i) follows by the bounded p.d.f of z_{1t} near zero and $T^{\alpha_{z,2}-\alpha_{x,2}} = o(1)$, as well as by Lemma 9.

Since $r > 8$, it is not hard to show that $r/3 > 2 + r/(2r-4) = 2 + \nu/2$. By Assumptions 1 and 2, $2\nu\alpha_{z,2} < \alpha_{x,2} < 1/2$. Since $r > 8$, we have that $r/3 > 8/3 > 2 + 1/2 > 2 + 2\nu\alpha_{z,2}$. Fix $p \in (2 + 2\nu\alpha_{z,2}, r/3)$. Now (71), (72) and Lemma 9 imply that for some constants $K_1, K_2 > 0$

$$\mathbb{E}\left(|\Delta L_{t+1,*}| \left|\mathbf{1}\{z_t'\tilde{\theta}_t > 0\} - \mathbf{1}\{z_{1t} > 0\}\right|\right)$$
$$\leq K_1 \|\Delta L_{t+1,*}\|_{L^q(\mathbb{P})} \left[T^{(\alpha_{z,2}-\alpha_{x,2})/\nu} + \left(T^{1-p/2}\log^{-K_2}\right)^{1/\nu}\right]. \tag{73}$$

By (66), we have that

$$\mathbb{E}\Delta L_{t+1,*}\mathbf{1}\{z_{1t} > 0\} \geq 2\mathbb{E}\left[\varepsilon_{t+1}\left(\beta_2 x_{2,t} - \beta_1 x_{1,t}\right)\mathbf{1}\{z_{1t} > 0\}\right] - A_t, \tag{74}$$

where

$$A_t = 2\mathbb{E}\left|\varepsilon_{t+1}\left(\delta_{2,t,*}x_{2,t} - \delta_{1,t,*}x_{1,t}\right)\right|$$
$$+ \mathbb{E}\left|\left(\beta_2 x_{2,t} - \beta_1 x_{1,t} - \delta_{1,t,*}x_{1,t} + \delta_{2,t,*}x_{2,t}\right)\left(\beta_1 x_{1,t} + \beta_2 x_{2,t} - \delta_{1,t,*}x_{1,t} - \delta_{2,t,*}x_{2,t}\right)\right|.$$

After computations similar to (67) in the proof of Lemma 8, we can use the rate conditions in Assumption 1 and show that for some constant $K_3 > 0$,

$$A_t \leq K_3 T^{-2\alpha_{x,2}}. \tag{75}$$

(74) and (75) imply that for some constants $K_4, K_5, K_6 > 0$, we have that for $T \geq K_4$,

$$\mathbb{E}\Delta L_{t+1,*}\mathbf{1}\{z_{1t} > 0\} \geq 2\mathbb{E}\left[\varepsilon_{t+1}\left(\beta_2 x_{2,t} - \beta_1 x_{1,t}\right)\mathbf{1}\{z_{1t} > 0\}\right] - K_3 T^{-1/2} - K_4 T^{-2\alpha_{x,2}}$$
$$\overset{(i)}{\geq} K_5 T^{-\alpha_{x,2}-\alpha_{z,2}} - K_3 T^{-2\alpha_{x,2}},$$

where (i) holds by Assumption 1. By the above display and (73), we have that for

large T,

$$\mathbb{E}S_{t+1} \geq K_5 T^{-\alpha_{x,2}-\alpha_{z,2}} - K_3 T^{-2\alpha_{x,2}} \\ - K_1 \|\Delta L_{t+1,*}\|_{L^q(\mathbb{P})} \left[T^{(\alpha_{z,2}-\alpha_{x,2})/\nu} + \left(T^{1-p/2} \log^{-K_2}\right)^{1/\nu} \right].$$

Recall that in Step 1 of the proof of Lemma 9, we have that $m^{\alpha_{x,2}} \Delta L_{t+1,*}$ has uniformly bounded $0.5r$-th moments. Since $q = 0.5r$, we have that $\|m^{\alpha_{x,2}} \Delta L_{t+1,*}\|_{L^q(\mathbb{P})}$ is bounded above by a constant. Hence, for some constant $K_7 > 0$,

$$\mathbb{E}S_{t+1} \geq K_5 T^{-\alpha_{x,2}-\alpha_{z,2}} - K_3 T^{-2\alpha_{x,2}} \\ - K_1 K_7 T^{-\alpha_{x,2}} \left[T^{(\alpha_{z,2}-\alpha_{x,2})/\nu} + \left(T^{1-p/2} \log^{-K_2}\right)^{1/\nu} \right]. \quad (76)$$

Since $p > 2 + 2\nu\alpha_{z,2}$ and $\nu = r/(r-2)$, it is not hard to show that $-\alpha_{x,2} + (1-p/2)/\nu < -\alpha_{x,2} - \alpha_{z,2}$. By Assumption 1, it is straight-forward to verify that $-2\alpha_{x,2} < -\alpha_{x,2} - \alpha_{z,2}$ and $-\alpha_{x,2} + (\alpha_{z,2} - \alpha_{x,2})/\nu < -\alpha_{x,2} - \alpha_{z,2}$. The desired result follows by (76). □

Proof of Proposition 2. We choose p_1 and h as in the proof of Proposition 1. Then Part (1) follows by Lemma 10 and the computations in the proof of Proposition 1. Part (2) follows by Part (1) and Proposition 1. □

B.3 Proof of Proposition 3

Our proof of Proposition 3 relies on two lemmas, Lemma 11 and 12. We first state and prove these lemmas before proving Proposition 3.

Lemma 11. *Suppose that the assumptions of Proposition 3 hold. Let $\Delta L_{t+1,*}$ be defined as in (66) in the proof of Lemma 8. Let $\tilde{\theta}_t = (\tilde{\theta}_{1,t}, \tilde{\theta}_{2,t})' = (\sum_{s=t-m}^{t-1} z_s z_s')^{-1} (\sum_{s=t-m}^{t-1} z_s \Delta L_{t+1,*})$, where $z_s = (1, z_{1s})'$. Fix $p \in (2, r/3)$. Then there exist some constants $G_0, G_1, G_2, G_3 > 0$ such that for $T \geq G_0$,*

$$\mathbb{P}\left(\left| \frac{\tilde{\theta}_{1,t}}{\tilde{\theta}_{2,t}} \right| \geq G_1 T^{\alpha_{x,1}+\alpha_{z,1}-2\alpha_{x,2}} \right) \leq G_2 T^{1-p/2} \log^{-G_3}$$

and

$$\mathbb{P}\left(\tilde{\theta}_{2,t} \leq 0 \right) \leq G_2 T^{1-p/2} \log^{-G_3}.$$

Proof. Let $\Psi_{t+1} = m^{\alpha_{x,2}} \Delta L_{t+1,*}$, $\pi_t = (\sum_{s=t-m}^{t-1} \mathbb{E} z_s z_s')^{-1} (\sum_{s=t-m}^{t-1} \mathbb{E} z_s \Psi_{s+1})$, $\hat{\pi}_t = (\sum_{s=t-m}^{t-1} z_s z_s')^{-1} (\sum_{s=t-m}^{t-1} z_s \Psi_{s+1})$ and $\{\xi_{s+1}\}_{s=t-m}^{t-1}$ with $\xi_{s+1} = \Psi_{s+1} - z_s' \pi_t$. Clearly, $\hat{\pi}_t = m^{\alpha_{x,2}} \tilde{\theta}_t$. Let $\gamma_t = \hat{\pi}_t - \pi_t$. The proof proceeds in two steps. We first bound γ_t and then show the desired results.

Step 1: bound γ_t

By simple computation, we have

$$\gamma_t = \underbrace{\left[m^{-1} \sum_{s=t-m}^{t-1} z_s z_s' \right]^{-1}}_{J_t} \cdot \underbrace{m^{-1} \sum_{s=t-m}^{t-1} z_s \xi_{s+1}}_{B_t}.$$

Notice that $p \in (2, r/3)$. Since entries of $z_s z_s' - \mathbb{E} z_s z_s'$ has uniformly bounded $0.5r$-th moments, it follows, by Lemma 4, that for some constants $K_1, K_2, K_3 > 0$,

$$\mathbb{P}\left(\left\| m^{-1} \sum_{s=t-m}^{t-1} (z_s z_s' - \mathbb{E} z_s z_s') \right\|_\infty \geq K_1 \sqrt{m^{-1} \log m} \right) \leq K_2 m^{1-p/2} \log^{-K_3} m.$$

Since $m^{-1} \sum_{s=t-m}^{t-1} \mathbb{E} z_s z_s' = \mathrm{diag}(1, m^{-1} \sum_{s=t-m}^{t-1} \mathbb{E} z_{1s}^2)$ and $\mathbb{E} z_{1s}^2$ is bounded away from zero and infinity, the eigenvalues of $m^{-1} \sum_{s=t-m}^{t-1} \mathbb{E} z_s z_s'$ lie in $[K_4, K_5]$ for some constants $K_4, K_5 > 0$. By the above display, the eigenvalues of $m^{-1} \sum_{s=t-m}^{t-1} z_s z_s'$ lie in $[K_4/2, 2K_5]$ for large n with probability at least $1 - K_2 m^{1-p/2} \log^{-K_3} m$. Hence, for some constants $K_6, K_7 > 0$, we have that for $n \geq K_6$,

$$\mathbb{P}\left(\|J_t\|_\infty \geq K_7 \right) \leq K_2 m^{1-p/2} \log^{-K_3} m.$$

Recall the definition of $\Delta L_{t+1,*}$ (in (66) in the proof of Lemma 8). We apply Lemma 2. By straight-forward computations, we have that Ψ_{t+1} has uniformly bounded $0.5r$-th moments. Hence, $\xi_{s+1} = \Psi_{s+1} - z_s' \pi_s$ has uniformly bounded $0.5r$-th moments. Again by Lemma 2, entries of $z_s \xi_{s+1}$ has uniformly bounded $\frac{r}{3}$-th moments. Notice that $p \in (2, r/3)$. It follows by 4 (applied to each entry of $z_s \xi_{s+1}$) that for some constants $K_8, K_9, K_{10} > 0$, we have

$$\mathbb{P}\left(\|B_t\|_\infty \geq K_8 \sqrt{m^{-1} \log m} \right) \leq K_9 m^{1-p/2} \log^{-K_{10}} m.$$

It follows by the above two displays that for large m,

$$\begin{aligned}\mathbb{P}\left(\|\gamma_t\|_\infty \geq 2K_7K_8\sqrt{m^{-1}\log m}\right) &\leq \mathbb{P}\left(2\|J_t\|_\infty\|B_t\|_\infty \geq 2K_7K_8\sqrt{m^{-1}\log m}\right)\\ &\leq \mathbb{P}\left(\|J_t\|_\infty \geq K_7\right) + \mathbb{P}\left(\|B_t\|_\infty \geq K_8\sqrt{m^{-1}\log m}\right)\\ &\leq K_2 m^{1-p/2}\log^{-K_3} m + K_9 m^{1-p/2}\log^{-K_{10}} m.\end{aligned}$$

Since $T \asymp m$, there are constants $K_{11}, ..., K_{14} > 0$ such that for $T \geq K_{11}$,

$$\mathbb{P}\left(\|\gamma_t\|_\infty \geq K_{12}\sqrt{T^{-1}\log T}\right) \leq K_{13}T^{1-p/2}\log^{-K_{14}}. \tag{77}$$

Step 2: show the desired results.

Partition $\pi_t = (\pi_{1,t}, \pi_{2,t})'$. By $\mathbb{E}z_{1s} = 0$, it follows that $\pi_{1,t} = m^{-1}\sum_{s=t-m}^{t-1}\mathbb{E}m^{\alpha_{x,2}}\Delta L_{s+1,*}$ and $\pi_{2,t} = (\sum_{s=t-m}^{t-1}\mathbb{E}m^{\alpha_{x,2}}z_{1s}\Delta L_{s+1,*})/(m^{-1}\sum_{s=t-m}^{t-1}\mathbb{E}z_{1s}^2)$. By Lemma 8, there are constants $K_{15}, K_{16} > 0$ such that

$$K_{15}T^{-\alpha_{x,2}} \leq \pi_{1,t} \leq K_{16}T^{-\alpha_{x,2}}. \tag{78}$$

By computations similar to (67) in the proof of Lemma 8, one can show that

$$\mathbb{E}\left|z_{1t}\Delta L_{t+1,*} - z_{1t}\left[2\varepsilon_{t+1}\left(\beta_2 x_{2,t} - \beta_1 x_{1,t}\right) + \beta_2^2 x_{2,t}^2 - \beta_{1,t}^2 x_{1,t}^2\right]\right| \leq K_{17}T^{-1/2}\sqrt{\log T},$$

where $K_{17} > 0$ is a constant. By the assumptions of Proposition 3,

$$\mathbb{E}\left\{z_{1t}\left[2\varepsilon_{t+1}\left(\beta_2 x_{2,t} - \beta_1 x_{1,t}\right) + \beta_2^2 x_{2,t}^2 - \beta_{1,t}^2 x_{1,t}^2\right]\right\} \geq K_{18}T^{-\alpha_{x,1}-\alpha_{z,1}}.$$

It follows that for large T, $\mathbb{E}z_{1t}\Delta L_{t+1,*} \geq K_{18}T^{-\alpha_{x,1}-\alpha_{z,1}}/2$. Hence, for some constant $K_{19} > 0$,

$$\pi_{2,t} \geq K_{19}T^{\alpha_{x,2}-\alpha_{x,1}-\alpha_{z,1}}. \tag{79}$$

Let $x = 2K_{16}T^{-\alpha_{x,2}}$ and $M = 2x/(K_{19}T^{\alpha_{x,2}-\alpha_{x,1}-\alpha_{z,1}})$. Then

$$\begin{aligned}&\mathbb{P}\left(|\hat{\pi}_{1,t}| \geq M|\hat{\pi}_{2,t}|\right)\\ &\leq \mathbb{P}\left(|\hat{\pi}_{1,t}| \geq x\right) + \mathbb{P}\left(|\hat{\pi}_{2,t}| \leq x/M\right)\\ &\leq \mathbb{P}\left(|\hat{\pi}_{1,t} - \pi_{1,t}| \geq x - |\pi_1|\right) + \mathbb{P}\left(|\hat{\pi}_{2,t} - \pi_{2,t}| \geq |\pi_{2,t}| - x/M\right)\end{aligned}$$

$$
\begin{aligned}
&\overset{(i)}{\leq} \mathbb{P}\left(|\hat{\pi}_{1,t} - \pi_{1,t}| \geq x - K_{16}T^{-\alpha_{x,2}}\right) + \mathbb{P}\left(|\hat{\pi}_{2,t} - \pi_{2,t}| \geq K_{19}T^{\alpha_{x,2}-\alpha_{x,1}-\alpha_{z,1}} - x/M\right) \\
&\leq \mathbb{P}\left(\|\gamma_t\|_\infty \geq x - K_{16}T^{-\alpha_{x,2}}\right) + \mathbb{P}\left(\|\gamma_t\|_\infty \geq K_{19}T^{\alpha_{x,2}-\alpha_{x,1}-\alpha_{z,1}} - x/M\right) \\
&= \mathbb{P}\left(\|\gamma_t\|_\infty \geq K_{16}T^{-\alpha_{x,2}}\right) + \mathbb{P}\left(\|\gamma_t\|_\infty \geq K_{19}T^{\alpha_{x,2}-\alpha_{x,1}-\alpha_{z,1}}/2\right) \\
&\overset{(ii)}{\leq} 2K_{13}T^{1-p/2}\log^{-K_{14}},
\end{aligned}
$$

where (i) holds by (78) and (79) and (ii) follows by (77) together with $T^{-\alpha_{x,2}} \gg \sqrt{T^{-1}\log T}$ and $T^{\alpha_{x,2}-\alpha_{x,1}-\alpha_{z,1}} \gg \sqrt{T^{-1}\log T}$. The first claim follows by $|\hat{\pi}_{1,t}/\hat{\pi}_{2,t}| = |\tilde{\theta}_{1,t}/\tilde{\theta}_{2,t}|$ and $M = 2x/(K_{19}T^{\alpha_{x,2}-\alpha_{x,1}-\alpha_{z,1}}) = (4K_{16}/K_{19})T^{\alpha_{x,1}+\alpha_{z,1}-2\alpha_{x,2}}$.

To see the second claim, notice that

$$
\begin{aligned}
\mathbb{P}\left(\tilde{\theta}_{2,t} \leq 0\right) \overset{(i)}{=} \mathbb{P}\left(\hat{\pi}_{2,t} \leq 0\right) = \mathbb{P}\left(\gamma_{2,t} \leq -\pi_{2,t}\right) &\overset{(ii)}{\leq} \mathbb{P}\left(\gamma_{2,t} \leq -K_{19}T^{\alpha_{x,2}-\alpha_{x,1}-\alpha_{z,1}}\right) \\
&\leq \mathbb{P}\left(\|\gamma_t\|_\infty \geq K_{19}T^{\alpha_{x,2}-\alpha_{x,1}-\alpha_{z,1}}\right) \\
&\overset{(iii)}{\leq} K_{13}T^{1-p/2}\log^{-K_{14}},
\end{aligned}
$$

where (i) holds by $\tilde{\theta}_t = m^{-\alpha_{x,2}}\hat{\pi}_t$, (ii) follows by (79) and (iii) holds by (77) and $T^{\alpha_{x,2}-\alpha_{x,1}-\alpha_{z,1}} \gg \sqrt{T^{-1}\log T}$. □

Lemma 12. *Let the assumptions of Proposition 3 hold. Then $\forall p_1 \in (2, r/3)$, there exist constants $G_0, G_1, ..., G_5 > 0$ and an array $\{S_{t+1}\}_{t=n+m}^{T-1}$ such that $\forall T \geq G_0$,*

$$
\mathbb{P}\left(\bigcap_{t=n+m}^{T-1}\left\{S_{t+1} = \Delta L_{t+1}\mathbf{1}\{z_t'\hat{\theta}_{m,t} > 0\}\right\}\right) \geq 1 - G_1T^{2-p_1/2}\log^{-G_2}T
$$

and

$$
\mathbb{E}S_{t+1} \geq K_5T^{-\alpha_{x,1}-\alpha_{z,1}}.
$$

Proof. Let $\tilde{\theta}_t = (\sum_{s=t-m}^{t-1} z_s z_s')^{-1}(\sum_{s=t-m}^{t-1} z_s \Delta L_{t+1,*})$, where $\Delta L_{t+1,*}$ is defined as in (66) in the proof of Lemma 8. Define $S_{t+1} = \Delta L_{t+1,*}\mathbf{1}\{z_t'\tilde{\theta}_t > 0\}$. Notice that

$$
\bigcap_{t=n}^{T-1}\{\Delta L_{t+1,*} = \Delta L_{t+1}\} \subseteq \bigcap_{t=n+m}^{T-1}\left\{S_{t+1} = \Delta L_{t+1}\mathbf{1}\{z_t'\hat{\theta}_{m,t} > 0\}\right\}.
$$

Hence, the first claim follows by Lemma 8.

It remains to bound $\mathbb{E}S_{t+1}$. To this end, let $q = r/2$ and $\nu = r/(r-2)$. Hence,

$q, \nu > 1$ and $q^{-1} + \nu^{-1} = 1$. Notice that

$$
\begin{aligned}
&\mathbb{E}\left(|\Delta L_{t+1,*}| \left|\mathbf{1}\{z_t'\tilde{\theta}_t > 0\} - \mathbf{1}\{z_{1t} > 0\}\right|\right) \\
&= \mathbb{E}\Bigg(|\Delta L_{t+1,*}| \cdot \Big|\mathbf{1}\{z_{1t} > -\tilde{\theta}_{1,t}/\tilde{\theta}_{2,t} \text{ and } \tilde{\theta}_{2,t} > 0\} \\
&\qquad\qquad + \mathbf{1}\{z_{1t} < -\tilde{\theta}_{1,t}/\tilde{\theta}_{2,t} \text{ and } \tilde{\theta}_{2,t} \leq 0\} - \mathbf{1}\{z_{1t} > 0\}\Big|\Bigg) \\
&\leq \mathbb{E}\left(|\Delta L_{t+1,*}| \left[\mathbf{1}\left\{0 < z_{1t} \leq -\tilde{\theta}_{1,t}/\tilde{\theta}_{2,t}\right\} + \mathbf{1}\left\{\tilde{\theta}_{2,t} \leq 0\right\}\right]\right) \\
&\overset{(i)}{\leq} \|\Delta L_{t+1,*}\|_{L^q(\mathbb{P})} \left\|\mathbf{1}\left\{0 < z_{1t} \leq -\tilde{\theta}_{1,t}/\tilde{\theta}_{2,t}\right\} + \mathbf{1}\left\{\tilde{\theta}_{2,t} \leq 0\right\}\right\|_{L^\nu(\mathbb{P})} \\
&\leq \|\Delta L_{t+1,*}\|_{L^q(\mathbb{P})} \left[\left\|\mathbf{1}\left\{0 < z_{1t} \leq -\tilde{\theta}_{1,t}/\tilde{\theta}_{2,t}\right\}\right\|_{L^\nu(\mathbb{P})} + \left\|\mathbf{1}\left\{\tilde{\theta}_{2,t} \leq 0\right\}\right\|_{L^\nu(\mathbb{P})}\right] \\
&= \|\Delta L_{t+1,*}\|_{L^q(\mathbb{P})} \left(\left[\mathbb{P}\left(0 < z_{1t} \leq -\tilde{\theta}_{1,t}/\tilde{\theta}_{2,t}\right)\right]^{1/\nu} + \left[\mathbb{P}\left(\tilde{\theta}_{2,t} \leq 0\right)\right]^{1/\nu}\right) \qquad (80)
\end{aligned}
$$

where (i) follows by Holder's inequality. By the assumptions of Proposition 3, the p.d.f of z_{1t} in a fixed neighborhood of zero is bounded above by some constant $K_0 > 0$. Recall constants $G_1, ..., G_4 > 0$ in the statement of Lemma 11. Hence,

$$
\begin{aligned}
&\mathbb{P}\left(0 < z_{1t} \leq -\tilde{\theta}_{1,t}/\tilde{\theta}_{2,t}\right) \\
&\leq \mathbb{P}\left(0 < z_{1t} \leq \left|\tilde{\theta}_{1,t}/\tilde{\theta}_{2,t}\right|\right) \\
&\leq \mathbb{P}\left(0 < z_{1t} \leq G_1 T^{\alpha_{x,1}+\alpha_{z,1}-2\alpha_{x,2}}\right) + \mathbb{P}\left(\left|\tilde{\theta}_{1,t}/\tilde{\theta}_{2,t}\right| \geq G_1 T^{\alpha_{x,1}+\alpha_{z,1}-2\alpha_{x,2}}\right) \\
&\overset{(i)}{\leq} K_0 G_1 T^{\alpha_{x,1}+\alpha_{z,1}-2\alpha_{x,2}} + G_2 T^{1-p/2} \log^{-G_3}, \qquad (81)
\end{aligned}
$$

where (i) follows by the bounded p.d.f of z_{1t} near zero and $T^{\alpha_{x,1}+\alpha_{z,1}-2\alpha_{x,2}} = o(1)$, as well as by Lemma 9.

It is not hard to show that $r/3 > 2 + r/(r-2) = 2 + \nu$ for $r \geq 10$. Fix $p \in (2 + \nu, r/3)$. Now (80), (81) and Lemma 11 imply that for some constants $K_1, K_2 > 0$

$$
\begin{aligned}
&\mathbb{E}\left(|\Delta L_{t+1,*}| \left|\mathbf{1}\{z_t'\tilde{\theta}_t > 0\} - \mathbf{1}\{z_{1t} > 0\}\right|\right) \\
&\qquad \leq K_1 \|\Delta L_{t+1,*}\|_{L^q(\mathbb{P})} \left[T^{(\alpha_{x,1}+\alpha_{z,1}-2\alpha_{x,2})/\nu} + \left(T^{1-p/2} \log^{-K_2}\right)^{1/\nu}\right]. \qquad (82)
\end{aligned}
$$

By (66), we have that

$$\mathbb{E}\Delta L_{t+1,*}\mathbf{1}\{z_{1t}>0\} \geq 2\mathbb{E}\left[\varepsilon_{t+1}\left(\beta_2 x_{2,t}-\beta_1 x_{1,t}\right)\mathbf{1}\{z_{1t}>0\}\right]-A_t, \tag{83}$$

where

$$\begin{aligned}A_t &= 2\mathbb{E}\left|\varepsilon_{t+1}\left(\delta_{2,t,*}x_{2,t}-\delta_{1,t,*}x_{1,t}\right)\right| \\ &+\mathbb{E}\left|\left(\beta_2 x_{2,t}-\beta_1 x_{1,t}-\delta_{1,t,*}x_{1,t}+\delta_{2,t,*}x_{2,t}\right)\left(\beta_1 x_{1,t}+\beta_2 x_{2,t}-\delta_{1,t,*}x_{1,t}-\delta_{2,t,*}x_{2,t}\right)\right|.\end{aligned}$$

After computations similar to (67) in the proof of Lemma 8, we can use the rate conditions in the assumptions of Proposition 3 and show that for some constant $K_3>0$,

$$A_t \leq K_3 T^{-2\alpha_{x,2}}. \tag{84}$$

(83) and (84) imply that for some constants $K_4, K_5, K_6 > 0$, we have that for $T \geq K_4$,

$$\begin{aligned}\mathbb{E}\Delta L_{t+1,*}\mathbf{1}\{z_{1t}>0\} &\geq 2\mathbb{E}\left[\varepsilon_{t+1}\left(\beta_2 x_{2,t}-\beta_1 x_{1,t}\right)\mathbf{1}\{z_{1t}>0\}\right]-K_3T^{-1/2}-K_4T^{-2\alpha_{x,2}} \\ &\overset{(i)}{\geq} K_5 T^{-\alpha_{x,1}-\alpha_{z,1}}-K_3T^{-2\alpha_{x,2}},\end{aligned}$$

where (i) holds by the assumptions of Proposition 3. By the above display and (82), we have that for large T,

$$\begin{aligned}\mathbb{E}S_{t+1} \geq\ & K_5 T^{-\alpha_{x,1}-\alpha_{z,1}}-K_3T^{-2\alpha_{x,2}} \\ & -K_1\|\Delta L_{t+1,*}\|_{L^q(\mathbb{P})}\left[T^{(\alpha_{x,1}+\alpha_{z,1}-2\alpha_{x,2})/\nu}+\left(T^{1-p/2}\log^{-K_2}\right)^{1/\nu}\right].\end{aligned}$$

Recall that in Step 1 of the proof of Lemma 9, we have that $m^{\alpha_{x,2}}\Delta L_{t+1,*}$ has uniformly bounded $0.5r$-th moments. Since $q=0.5r$, we have that $\|m^{\alpha_{x,2}}\Delta L_{t+1,*}\|_{L^q(\mathbb{P})}$ is bounded above by some constant $K_7>0$. Hence,

$$\begin{aligned}\mathbb{E}S_{t+1} \geq\ & K_5 T^{-\alpha_{x,1}-\alpha_{z,1}}-K_3T^{-2\alpha_{x,2}} \\ & -K_1K_7T^{-\alpha_{x,2}}\left[T^{(\alpha_{x,1}+\alpha_{z,1}-2\alpha_{x,2})/\nu}+\left(T^{1-p/2}\log^{-K_2}\right)^{1/\nu}\right].\end{aligned} \tag{85}$$

Since $p>\nu+2$ and $\nu=r/(r-2)$, it is not hard to show that $-\alpha_{x,2}+(1-$

$p/2)/\nu < (1-p/2)/\nu < -1/2 < -\alpha_{x,1} - \alpha_{z,1}$. By the assumptions of Proposition 3, $-2\alpha_{x,2} < -\alpha_{x,1} - \alpha_{z,1}$ and $-\alpha_{x,2} + (\alpha_{x,1} + \alpha_{z,1} - 2\alpha_{x,2})/\nu < -\alpha_{x,1} - \alpha_{z,1}$. The desired result follows by (85). □

Proof of Proposition 3. Part (1) follows by Lemma 12 and the arguments in the proof of Proposition 1. Part (2) follows by Part (1) and Proposition 1. □

B.4 Proof of Proposition 4

Our proof of Proposition 4 relies on three lemmas, lemmas 13-15. We first state and prove these lemmas before proving Proposition 4.

Lemma 13. *Let Assumption 4 hold. Define* $\Delta_{t,big} = (\hat{\mu}_t - \mu, \hat{\beta}_t - \beta)'$. *For any constants* $h \in (0,1)$ *and* $p_1 \in (2, r/2)$, *we can enlarge the probability space and construct random variables* $\delta_{t,small,*}$, $\Delta_{t,big,*}$, $\bar{\delta}_{t,small}$ *and* $\bar{\Delta}_{t,big}$ *such that*

$$\begin{cases} \mathbb{P}\left(\delta_{t,small,*} \neq \delta_{t,small}\right) \leq C_1 T^{-\min\{p_1/2-1,\ (1-h)p_1\}} \\ \mathbb{P}\left(|\delta_{t,small,*} - \bar{\delta}_{t,small}| \leq T^{-h}\right) = 1 \\ \bar{\delta}_{t,small} \text{ is independent of } \{x_s, \varepsilon_{s+1}\}_{s \geq t-1} \\ |\mathbb{E}\bar{\delta}_{t,small}| \leq T^{-1}\sqrt{\log T} C_2 \text{ and } T^{-1}C_3 \leq \mathbb{E}\bar{\delta}^2_{t,small} \leq T^{-1}C_4 \\ \left|\mathbb{E}\bar{\delta}^2_{t,small} - n^{-2}\mathbb{E}(\sum_{s=t-n}^{t-1} x_s\beta + \varepsilon_{s+1})^2\right| \leq C_5\sqrt{T^{-3}\log T} \end{cases}$$

and

$$\begin{cases} \mathbb{P}\left(\Delta_{t,big,*} \neq \Delta_{t,big}\right) \leq C_1 T^{-\min\{p_1/2-1,\ (1-h)p_1\}} \\ \mathbb{P}\left(\|\Delta_{t,big,*} - \bar{\Delta}_{t,big}\|_\infty \leq T^{-h}\right) = 1 \\ \bar{\Delta}_{t,big} \text{ is independent of } \{x_s, \varepsilon_{s+1}\}_{s \geq t-1} \\ \|\mathbb{E}\bar{\Delta}_{t,big}\|_\infty \leq T^{-1}\sqrt{\log T} C_2 \text{ and } T^{-1}C_3 \leq \mathbb{E}\|\bar{\Delta}_{t,big}\|^2_\infty \leq T^{-1}C_4 \\ \left\|\mathbb{E}\bar{\Delta}_{t,big}\bar{\Delta}'_{t,big} - \Sigma^{-1}_{X,t}\mathbb{E}(n^{-2}\sum_{s=t-n}^{t-1} \bar{x}_s\bar{x}'_s\varepsilon^2_{s+1})\Sigma^{-1}_{X,t}\right\|_\infty \leq C_5\sqrt{T^{-3}\log T}, \end{cases}$$

where $\bar{x}_t = (1, x_t)'$, $\Sigma_{X,t} = n^{-1}\sum_{s=t-n}^{t-1} \bar{x}_s\bar{x}'_s$ *and* $C_1, ..., C_4 > 0$ *are constants depending only on the constants in Assumption 4.*

Proof. The result follows by essentially the same argument as in the proof of Lemma 7. For results on $\Delta_{t,big}$, adjustments to allow multivariate $\bar{x}_s$ are needed but the arguments are essentially identical. □

Lemma 14. *Let Assumption 4 hold. Let $\alpha_x < 1/2$. Then $\forall p_1 \in (2, r/2)$ and $\forall h \in (0,1)$, there exist constants $G_1, ..., G_4 > 0$ and an array of random variables $\{\Delta L_{t+1,*}\}_{t=n}^{T-n}$ such that for $T \geq G_1$,*

$$\mathbb{P}\left(\bigcap_{t=n}^{T-1} \{\Delta L_{t+1,*} = \Delta L_{t+1}\}\right) \geq 1 - G_2 T^{1-\min\{p_1/2-1,\ (1-h)p_1\}}$$

and

$$G_3 T^{-2\alpha_x} \leq \mathbb{E}\Delta L_{t+1,*} \leq G_4 T^{-2\alpha_x}.$$

Proof. Recall from (17) that

$$\Delta L_{t+1} = \beta^2 x_t^2 + 2\beta x_t \varepsilon_{t+1} + \delta_{t,small}^2 - \delta_{t,big}^2 + 2\delta_{t,big}\varepsilon_{t+1} - 2\delta_{t,small}(\beta x_t + \varepsilon_{t+1}). \quad (86)$$

Recall $\delta_{t,small} = \tilde{\mu}_t - \mu$. Let $\Delta_{t,big} = (\hat{\mu}_t - \mu, \hat{\beta}_t - \beta)'$. Let $\delta_{t,small,*}$, $\Delta_{t,big,*}$, $\bar{\delta}_{t,small}$ and $\bar{\Delta}_{t,big}$ be defined as in Lemma 13. We define $\delta_{t,big,*} = \Delta'_{t,big,*}\bar{x}_t$ with $\bar{x}_t = (1, x_t)'$ and

$$\Delta L_{t+1,*} = \beta^2 x_t^2 + 2\beta x_t \varepsilon_{t+1} + \delta_{t,small,*}^2 - \delta_{t,big,*}^2 + 2\delta_{t,big,*}\varepsilon_{t+1} - 2\delta_{t,small,*}(\beta x_t + \varepsilon_{t+1}). \quad (87)$$

The first claim follows by Lemma 13.

By computations similar to (67) using Lemma 13, we have that

$$\mathbb{E}|\Delta L_{t+1,*} - \beta^2 x_t^2| \leq K T^{-h_1}$$

for some constants $K > 0$ and $h_1 \in (2\alpha_x, 1)$. The second claim follows. □

Lemma 15. *Let Assumption 4 hold. Let $\alpha_x > 1/2$. Then $\forall p_1 \in (2, r/2)$ and $\forall h \in (0,1)$, there exist constants $G_1, ..., G_4 > 0$ and an array of random variables $\{\Delta L_{t+1,*}\}_{t=n}^{T-n}$ such that for $T \geq G_1$,*

$$\mathbb{P}\left(\bigcap_{t=n}^{T-1} \{\Delta L_{t+1,*} = \Delta L_{t+1}\}\right) \geq 1 - G_2 T^{1-\min\{p_1/2-1,\ (1-h)p_1\}}$$

and

$$G_3 T^{-1} \leq \mathbb{E}\Delta L_{t+1,*} \leq G_4 T^{-1}.$$

Proof. Consider $\Delta L_{t+1,*}$ defined in (87). Recall $\delta_{t,small,*}$, $\Delta_{t,big,*}$, $\bar{\delta}_{t,small}$ and $\bar{\Delta}_{t,big}$ be

defined as in Lemma 13. Define

$$\overline{\Delta L}_{t+1} = \beta^2 x_t^2 + 2\beta x_t \varepsilon_{t+1} + \bar{\delta}^2_{t,small} - \bar{\delta}^2_{t,big} + 2\bar{\delta}_{t,big}\varepsilon_{t+1} - 2\bar{\delta}_{t,small}(\beta x_t + \varepsilon_{t+1}).$$

By computations similar to (67) using Lemma 13, we have that

$$\mathbb{E}|\Delta L_{t+1,*} - \overline{\Delta L}_{t+1}| \leq K T^{-1/2-\alpha_x} \tag{88}$$

for some constant $K > 0$. Let $\bar{x}_t = (1, x_t)'$. Now we compute

$$\begin{aligned}\mathbb{E}\overline{\Delta L}_{t+1} &= \beta^2 \mathbb{E}x_t^2 + \mathbb{E}\bar{\delta}^2_{t,small} - \mathbb{E}\left(\bar{x}_t'\bar{\Delta}_{t,big}\right)^2 + 2\mathbb{E}\bar{\Delta}'_{t,big}\bar{x}_t\varepsilon_{t+1} - 2\mathbb{E}\bar{\delta}_{t,small}(\beta x_t + \varepsilon_{t+1}) \\ &\overset{(i)}{=} \beta^2 \mathbb{E}x_t^2 + \mathbb{E}\bar{\delta}^2_{t,small} - \mathbb{E}\left(\bar{x}_t'\bar{\Delta}_{t,big}\bar{\Delta}'_{t,big}\bar{x}_t\right),\end{aligned}$$

where (i) follows by the fact that $\bar{\Delta}_{t,big}$ and $\bar{\delta}_{t,small}$ are independent of $\bar{x}_t$ and ε_{t+1}. By Lemma 13,

$$\begin{aligned}&\left|\left[\mathbb{E}\bar{\delta}^2_{t,small} - \mathbb{E}\left(\bar{x}_t'\bar{\Delta}_{t,big}\right)^2\right] - \left[n^{-2}\mathbb{E}(\sum_{s=t-n}^{t-1} x_s\beta + \varepsilon_{s+1})^2 - \mathbb{E}(\bar{x}_t'\Sigma_{X,t}^{-1}\Omega_t\Sigma_{X,t}^{-1}\bar{x}_t)\right]\right| \\ &\qquad\qquad \leq K_1\sqrt{T^{-3}\log T}\end{aligned}$$

for some constant $K_1 > 0$, where $\Sigma_{X,t} = n^{-1}\sum_{s=t-n}^{t-1}\mathbb{E}\bar{x}_s\bar{x}_s'$ and $\Omega_t = \mathbb{E}(n^{-2}\sum_{s=t-n}^{t-1}\bar{x}_s\bar{x}_s'\varepsilon_{s+1}^2)$. Notice that $\Sigma_{X,t} = \mathrm{diag}(1, \sigma_{x,t}^2)$ with $\sigma_{x,t}^2 = n^{-1}\sum_{s=t-n}^{t-1}\mathbb{E}x_s^2$. Hence,

$$\begin{aligned}&\mathbb{E}(\bar{x}_t'\Sigma_{X,t}^{-1}\Omega_t\Sigma_{X,t}^{-1}\bar{x}_t) \\ &\qquad = n^{-2}\sum_{s=t-n}^{t-1}\mathbb{E}\varepsilon_{s+1}^2 + n^{-2}\sum_{s=t-n}^{t-1}\sigma_{x,t}^{-4}\mathbb{E}(x_t^2x_s^2\varepsilon_{s+1}^2) + 2n^{-2}\sum_{s=t-n}^{t-1}\sigma_{x,t}^{-2}\mathbb{E}x_tx_s\varepsilon_{s+1}^2.\end{aligned}$$

It follows that

$$\begin{aligned}&n^{-2}\mathbb{E}\left(\sum_{s=t-n}^{t-1} x_s\beta + \varepsilon_{s+1}\right)^2 - \mathbb{E}(\bar{x}_t'\Sigma_{X,t}^{-1}\Omega_t\Sigma_{X,t}^{-1}\bar{x}_t) \\ &\qquad = n^{-1}\beta^2\sigma_{x,t}^2 - n^{-2}\sum_{s=t-n}^{t-1}\sigma_{x,t}^{-4}\mathbb{E}(x_t^2x_s^2\varepsilon_{s+1}^2) - 2n^{-2}\sum_{s=t-n}^{t-1}\sigma_{x,t}^{-2}\mathbb{E}x_tx_s\varepsilon_{s+1}^2.\end{aligned}$$

By computations based on the exponential decay of β-mixing coefficients similar to (51) and (52), it is not hard to show that $\sum_{s=t-n}^{t-1} \sigma_{x,t}^{-2}\mathbb{E}x_t x_s \varepsilon_{s+1}^2$ is uniformly bounded by a constant $K_2 > 0$. Hence,

$$\left|\mathbb{E}\overline{\Delta L}_{t+1} - \beta^2 \mathbb{E}x_t^2 - n^{-1}\beta^2\sigma_{x,t}^2 + n^{-2}\sum_{s=t-n}^{t-1} \sigma_{x,t}^{-4}\mathbb{E}(x_t^2 x_s^2 \varepsilon_{s+1}^2)\right| \leq 2n^{-2}K_2 + K_1\sqrt{T^{-3}\log T}.$$

Since $\sigma_{x,t}^{-4}\mathbb{E}(x_t^2 x_s^2 \varepsilon_{s+1}^2)$ and $\sigma_{x,t}^2$ are bounded away from zero and infinity, $\beta \asymp T^{-\alpha_x}$ with $\alpha_x > 1/2$, it follows that $-\mathbb{E}\overline{\Delta L}_{t+1} \asymp T^{-1}$. The desired result follows by (88) and $\alpha_x > 1/2$. □

Proof of Proposition 4. Part (1) follows by Lemma 14 and the arguments in the proof of Proposition 1. Part (2) follows by Lemma 15 and the arguments in the proof of Proposition 1. □

B.5 Proof of Proposition 5

Our proof of Proposition 5 relies on two lemmas, Lemmas 16 and 17. We first state and prove these lemmas before proving Proposition 5.

Lemma 16. *Suppose that the assumptions of Proposition 5 hold. Let $\Delta L_{t+1,*}$ be defined as in (66) in the proof of Lemma 8. Let $\tilde{\theta}_t = (\tilde{\theta}_{1,t}, \tilde{\theta}_{2,t})' = (\sum_{s=t-m}^{t-1} z_s z_s')^{-1}(\sum_{s=t-m}^{t-1} z_s \Delta L_{t+1,*})$, where $z_s = (1, z_{1s})'$. Fix $p \in (2, r/3)$. Then there exist some constants $G_0, G_1, G_2, G_3 > 0$ such that for $T \geq G_0$,*

$$\mathbb{P}\left(\left|\frac{\tilde{\theta}_{1,t}}{\tilde{\theta}_{2,t}}\right| \geq G_1 T^{\alpha_z - \alpha_x}\right) \leq G_2 T^{1-p/2}\log^{-G_3}$$

and

$$\mathbb{P}\left(\tilde{\theta}_{2,t} \leq 0\right) \leq G_2 T^{1-p/2}\log^{-G_3}.$$

Proof. The proof is similar to the proof of Lemma 11. Let $\Psi_{t+1} = m^{\alpha_x}\Delta L_{t+1,*}$, $\pi_t = (\sum_{s=t-m}^{t-1}\mathbb{E}z_s z_s')^{-1}(\sum_{s=t-m}^{t-1}\mathbb{E}z_s\Psi_{s+1})$, $\hat{\pi}_t = (\sum_{s=t-m}^{t-1} z_s z_s')^{-1}(\sum_{s=t-m}^{t-1} z_s\Psi_{s+1})$ and $\{\xi_{s+1}\}_{s=t-m}^{t-1}$ with $\xi_{s+1} = \Psi_{s+1} - z_s'\pi_t$. Clearly, $\hat{\pi}_t = m^{\alpha_x}\tilde{\theta}_t$. Let $\gamma_t = \hat{\pi}_t - \pi_t$. By the same argument as Step 1 in the proof of Lemma 11, we can show that there exist constants $M_1, M_2, M_3 > 0$ such that

$$\mathbb{P}\left(\|\gamma_t\|_\infty \geq M_1\sqrt{T^{-1}\log T}\right) \leq M_2 T^{1-p/2}\log^{-M_3}. \tag{89}$$

Now we characterize π_t. By Lemma 8, there are constants $M_4, M_5 > 0$ such that

$$M_4 T^{-\alpha_x} \leq \pi_{1,t} \leq M_5 T^{-\alpha_x}. \tag{90}$$

By computations similar to (67) in the proof of Lemma 8, one can show that

$$\mathbb{E}\left|z_{1t}\Delta L_{t+1,*} - z_{1t}(2\beta x_t \varepsilon_{t+1} + \beta^2 x_t^2)\right| \leq M_6 T^{-1/2}\sqrt{\log T},$$

where $M_6 > 0$ is a constant. By the assumptions of Proposition 5, there exists a constant $M_7 > 0$ with

$$\mathbb{E}\left[z_{1t}\left(2\beta x_t \varepsilon_{t+1} + \beta^2 x_t^2\right)\right] \geq M_7 T^{-\alpha_x - \alpha_z}.$$

Hence, for some constant $M_8 > 0$ we have

$$\pi_{2,t} \geq M_8 T^{-\alpha_z}. \tag{91}$$

Let $x = 2M_5 T^{-\alpha_x}$ and $G = 2x/(M_8 T^{-\alpha_z})$. Then

$$\begin{aligned}
&\mathbb{P}\left(|\hat{\pi}_{1,t}| \geq G|\hat{\pi}_{2,t}|\right) \\
&\leq \mathbb{P}\left(|\hat{\pi}_{1,t}| \geq x\right) + \mathbb{P}\left(|\hat{\pi}_{2,t}| \leq x/G\right) \\
&\leq \mathbb{P}\left(|\hat{\pi}_{1,t} - \pi_{1,t}| \geq x - |\pi_1|\right) + \mathbb{P}\left(|\hat{\pi}_{2,t} - \pi_{2,t}| \geq |\pi_{2,t}| - x/G\right) \\
&\overset{(i)}{\leq} \mathbb{P}\left(|\hat{\pi}_{1,t} - \pi_{1,t}| \geq x - M_5 T^{-\alpha_x}\right) + \mathbb{P}\left(|\hat{\pi}_{2,t} - \pi_{2,t}| \geq M_8 T^{-\alpha_z} - x/G\right) \\
&\leq \mathbb{P}\left(\|\gamma_t\|_\infty \geq x - M_5 T^{-\alpha_x}\right) + \mathbb{P}\left(\|\gamma_t\|_\infty \geq M_8 T^{-\alpha_z} - x/G\right) \\
&= \mathbb{P}\left(\|\gamma_t\|_\infty \geq M_5 T^{-\alpha_{x,2}}\right) + \mathbb{P}\left(\|\gamma_t\|_\infty \geq M_8 T^{-\alpha_z}/2\right) \\
&\overset{(ii)}{\leq} 2M_2 T^{1-p/2}\log^{-M_3},
\end{aligned}$$

where (i) holds by (90) and (91) and (ii) follows by (89) together with $T^{-\alpha_x} \gg \sqrt{T^{-1}\log T}$ and $T^{-\alpha_z} \gg \sqrt{T^{-1}\log T}$. The first claim follows by $|\hat{\pi}_{1,t}/\hat{\pi}_{2,t}| = |\tilde{\theta}_{1,t}/\tilde{\theta}_{2,t}|$ and $G = 2x/(M_8 T^{-\alpha_z}) = (4M_5/M_8)T^{\alpha_z - \alpha_x}$.

To see the second claim, notice that

$$\mathbb{P}\left(\tilde{\theta}_{2,t} \leq 0\right) \overset{(i)}{=} \mathbb{P}\left(\hat{\pi}_{2,t} \leq 0\right) = \mathbb{P}\left(\gamma_{2,t} \leq -\pi_{2,t}\right)$$

$$\overset{(ii)}{\leq} \mathbb{P}\left(\gamma_{2,t} \leq -M_8 T^{-\alpha_z}\right) \leq \mathbb{P}\left(\|\gamma_t\|_\infty \geq M_8 T^{-\alpha_z}\right)$$
$$\overset{(iii)}{\leq} M_2 T^{1-p/2} \log^{-M_3} ,$$

where (i) holds by $\tilde{\theta}_t = m^{-\alpha_x}\hat{\pi}_t$, (ii) follows by (79) and (iii) holds by (89) and $T^{-\alpha_z} \gg \sqrt{T^{-1}\log T}$. □

Lemma 17. *Let the assumptions of Proposition 5 hold. Then $\forall p_1 \in (2, r/3)$, there exist constants $G_0, G_1, ..., G_5 > 0$ and an array $\{S_{t+1}\}_{t=n+m}^{T-1}$ such that $\forall T \geq G_0$,*

$$\mathbb{P}\left(\bigcap_{t=n+m}^{T-1} \left\{S_{t+1} = \Delta L_{t+1} \mathbf{1}\{z_t'\hat{\theta}_{m,t} > 0\}\right\}\right) \geq 1 - G_1 T^{2-p_1/2}\log^{-G_2} T$$

and

$$\mathbb{E}S_{t+1} \geq K_5 T^{-\alpha_x-\alpha_z}.$$

Proof. The proof is similar to the proof of Lemma 12. Let $\tilde{\theta}_t = (\sum_{s=t-m}^{t-1} z_s z_s')^{-1}(\sum_{s=t-m}^{t-1} z_s \Delta L_{t+1,*})$, where $\Delta L_{t+1,*}$ is defined as in (87). Define $S_{t+1} = \Delta L_{t+1,*}\mathbf{1}\{z_t'\tilde{\theta}_t > 0\}$. Notice that

$$\bigcap_{t=n}^{T-1} \{\Delta L_{t+1,*} = \Delta L_{t+1}\} \subseteq \bigcap_{t=n+m}^{T-1} \left\{S_{t+1} = \Delta L_{t+1}\mathbf{1}\{z_t'\hat{\theta}_{m,t} > 0\}\right\}.$$

Hence, the first claim follows by Lemma 14.

To show the second claim, let $q = (r+2)/4$ and $\nu = (r+2)/(r-2)$. Hence, $q^{-1} + \nu^{-1} = 1$. Notice that by the same argument as (80) in the proof of Lemma 12, we have that

$$\mathbb{E}\left(|\Delta L_{t+1,*}| \left|\mathbf{1}\{z_t'\tilde{\theta}_t > 0\} - \mathbf{1}\{z_{1t} > 0\}\right|\right)$$
$$\leq \|\Delta L_{t+1,*}\|_{L^q(\mathbb{P})} \left(\left[\mathbb{P}\left(0 < z_{1t} \leq -\tilde{\theta}_{1,t}/\tilde{\theta}_{2,t}\right)\right]^{1/\nu} + \left[\mathbb{P}\left(\tilde{\theta}_{2,t} \leq 0\right)\right]^{1/\nu}\right).$$

Similar to the argument in (81), we have

$$\mathbb{P}\left(0 < z_{1t} \leq -\tilde{\theta}_{1,t}/\tilde{\theta}_{2,t}\right) \leq \mathbb{P}\left(0 < z_{1t} \leq \left|\tilde{\theta}_{1,t}/\tilde{\theta}_{2,t}\right|\right)$$
$$\leq \mathbb{P}\left(0 < z_{1t} \leq G_1 T^{\alpha_z-\alpha_x}\right) + \mathbb{P}\left(\left|\tilde{\theta}_{1,t}/\tilde{\theta}_{2,t}\right| \geq G_1 T^{\alpha_z-\alpha_x}\right)$$

$$\overset{(i)}{\leq} K_1 G_1 T^{\alpha_z - \alpha_x} + G_2 T^{1-p/2} \log^{-G_3}$$

for some constant $K_1 > 0$, where (i) holds by the bounded p.d.f of z_{1t} around zero and Lemma 16. The above two displays and Lemma 16 imply that for some constant $K_2 > 0$,

$$\mathbb{E}\left(|\Delta L_{t+1,*}| \left|\mathbf{1}\{z_t'\tilde{\theta}_t > 0\} - \mathbf{1}\{z_{1t} > 0\}\right|\right) \leq K_2 \|\Delta L_{t+1,*}\|_{L^q(\mathbb{P})} \left(T^{(\alpha_z - \alpha_x)/\nu} + (T^{1-p/2}\log^{-G_3})^{1/\nu}\right).$$

By (66), we have that

$$\mathbb{E}\Delta L_{t+1,*}\mathbf{1}\{z_{1t} > 0\} \geq 2\mathbb{E}\left[\beta x_t \varepsilon_{t+1}\mathbf{1}\{z_{1t} > 0\}\right] - A_t,$$

where

$$A_t = \mathbb{E}\left|\beta^2 x_t^2 + \delta_{t,small,*}^2 - \delta_{t,big,*}^2 + 2\delta_{t,big}\varepsilon_{t+1} - 2\delta_{t,small,*}(\beta x_t + \varepsilon_{t+1})\right|.$$

After computations similar to (67) in the proof of Lemma 8, we can use the rate conditions in the assumptions of Proposition 5 and show that for some constant $K_3 > 0$,

$$A_t \leq K_3(T^{-2\alpha_x} + T^{-1/2}). \tag{92}$$

(92) implies that for some constants $K_4, K_5 > 0$, we have that for $T \geq K_4$,

$$\begin{aligned}\mathbb{E}\Delta L_{t+1,*}\mathbf{1}\{z_{1t} > 0\} &\geq 2\mathbb{E}\left[\beta x_t \varepsilon_{t+1}\mathbf{1}\{z_{1t} > 0\}\right] - K_3(T^{-2\alpha_x} + T^{-1/2}) \\ &\overset{(i)}{\geq} K_5 T^{-\alpha_x - \alpha_z} - K_3(T^{-2\alpha_x} + T^{-1/2}),\end{aligned}$$

where (i) holds by the assumptions of Proposition 5. By the above display and (82), we have that for $T \geq K_4$,

$$\begin{aligned}\mathbb{E}S_{t+1} \geq\ & K_5 T^{-\alpha_x - \alpha_z} - K_3(T^{-2\alpha_x} + T^{-1/2}) \\ & - K_2 \|\Delta L_{t+1,*}\|_{L^q(\mathbb{P})} \left(T^{(\alpha_z - \alpha_x)/\nu} + (T^{1-p/2}\log^{-G_3})^{1/\nu}\right).\end{aligned}$$

Recall that in Step 1 of the proof of Lemma 9, we have that $T^{\alpha_x}\Delta L_{t+1,*}$ has uniformly bounded $0.5r$-th moments. Since $q = (r+2)/4 < 0.5r$, we have that

$\|T^{\alpha_x}\Delta L_{t+1,*}\|_{L^q(\mathbb{P})}$ is bounded above by some constant $K_6 > 0$. Hence,

$$\mathbb{E}S_{t+1} \geq K_5 T^{-\alpha_x-\alpha_z} - K_3(T^{-2\alpha_x} + T^{-1/2}) \\ - K_2K_6T^{-\alpha_x}\left(T^{(\alpha_z-\alpha_x)/\nu} + (T^{1-p/2}\log^{-G_3})^{1/\nu}\right).$$

It is not hard to show that $-\alpha_x - \alpha_z > -2\alpha_x$, $-\alpha_x - \alpha_z > -1/2$, $-\alpha_x - \alpha_z > -\alpha_x + (\alpha_z - \alpha_x)/\nu$ and $-\alpha_x - \alpha_z > -\alpha_x + (1-p/2)/\nu$. The desired result follows. □

Proof of Proposition 5. Part (1) follows by Lemma 17 and the arguments in the proof of Proposition 1. Part (2) follows by Part (1) and Proposition 4. □

B.6 Proof of Proposition 8

Our proof of Proposition 8 relies on two lemmas, Lemmas 18 and 19. We first state and prove these lemmas before proving Proposition 8.

Lemma 18. *Let the assumptions of Proposition 8 hold. Then there exist constants $K_1, K_2, K_3 > 0$ such that for any $G \geq 1$,*

$$\mathbb{E}\left|\mathbb{E}(x_t^2 \mid z_{1t}) - \mathbb{E}x_t^2\right| \leq K_1 G + K_2\sqrt{G(t-G)} + K_3[\alpha(G)]^{C_0/(C_0+2)}(t-G).$$

Proof. Notice that $x_t = \phi^G x_{t-G} + D_t$, where $D_t = \sum_{k=t-G+1}^{t}\phi^{t-k}u_k$. The proof proceeds in two steps. First, we derive a bound for $|\mathbb{E}x_t^2 - \phi^{2G}\mathbb{E}x_{t-G}^2|$ and $\mathbb{E}\left|\mathbb{E}(x_t^2 \mid z_{1t}) - \phi^{2G}\mathbb{E}(x_{t-G}^2 \mid z_{1t})\right|$; then we derive a bound for $\mathbb{E}|\mathbb{E}(x_{t-G}^2 \mid z_{1t}) - \mathbb{E}x_{t-G}^2|$.

Step 1: bound $|\mathbb{E}x_t^2 - \phi^{2G}\mathbb{E}x_{t-G}^2|$ and $\left|\mathbb{E}(x_t^2 \mid z_{1t}) - \phi^{2G}\mathbb{E}(x_{t-G}^2 \mid z_{1t})\right|$.

Clearly, $x_t^2 - \phi^{2G}x_{t-G}^2 = D_t^2 + 2\phi^{2G}x_{t-G}D_t$. Let $M_1 > 1$ be a constant satisfying $\mathbb{E}u_t^2 \leq M_1$ and $\mathbb{E}|u_t|^{2+C_0} \leq M_1$. Notice that

$$\mathbb{E}D_t^2 = \sum_{k=t-G+1}^{t}\phi^{2(t-k)}\mathbb{E}u_k^2 \leq GM_1. \tag{93}$$

Also notice that x_{t-G} and D_t are independent and can be written as $x_{t-G} = \sum_{k=1}^{t-G}\phi^{t-G-k}u_k$ and $D_t = \sum_{k=t-G+1}^{t}\phi^{t-k}u_k$. Notice that by Rosenthal's inequalities (Theorem 9.1 in Gut (2013)), there exists some constant M_2 depending only C_0 such

that

$$\mathbb{E}|x_{t-G}|^{2+C_0} \leq M_2 \max\left\{\sum_{k=1}^{t-G}\phi^{(t-G-k)(2+C_0)}\mathbb{E}|u_k|^{2+C_0},\ \left(\sum_{k=1}^{t-G}\phi^{2(t-G-k)}\mathbb{E}u_k^2\right)^{(1+C_0/2)}\right\}$$
$$\overset{(i)}{\leq} M_2 M_1^{1+C_0/2}(t-G)^{1+C_0/2}, \tag{94}$$

where (i) holds by $|\phi| \leq 1$, $\mathbb{E}u_k^2 \leq M_1$ and $\mathbb{E}|u_k|^{2+C_0} \leq M_1$. Similarly, we have $\mathbb{E}|D_t|^{2+C_0} \leq M_2 M_1^{1+C_0/2} G^{1+C_0/2}$. Therefore,

$$\mathbb{E}|x_{t-G}D_t| \overset{(i)}{=} \mathbb{E}|x_{t-G}| \cdot \mathbb{E}|D_t| \leq \left(\mathbb{E}|x_{t-G}|^{2+C_0}\right)^{1/(2+C_0)}\left(\mathbb{E}|D_t|^{2+C_0}\right)^{1/(2+C_0)}$$
$$\overset{(i)}{\leq} M_1 M_2^{2/(C_0+2)}\sqrt{G(t-G)}, \tag{95}$$

where (i) holds by the independence between X_{t-G} and D_t and (ii) follows by the bounds of $\mathbb{E}|x_{t-G}|^{2+C_0}$ and $\mathbb{E}|D_t|^{2+C_0}$.

By (93), (95) and $|\phi| \leq 1$, we have that $\mathbb{E}D_t^2 + 2\phi^{2G}\mathbb{E}|x_{t-G}D_t| \leq GM_1 + 2M_1 M_2^{2/(C_0+2)}\sqrt{G(t-G)}$. Since both $|\mathbb{E}x_t^2 - \phi^{2G}\mathbb{E}x_{t-G}^2|$ and $\mathbb{E}\left|\mathbb{E}(x_t^2 - \phi^{2G}x_{t-G}^2 \mid z_{1t})\right|$ is bounded above by $\mathbb{E}|x_t^2 - \phi^{2G}x_{t-G}^2|$, we have that

$$\begin{aligned}
&\max\left\{|\mathbb{E}x_t^2 - \phi^{2G}\mathbb{E}x_{t-G}^2|,\ \mathbb{E}\left|\mathbb{E}(x_t^2 - \phi^{2G}x_{t-G}^2 \mid z_{1t})\right|\right\}\\
&\leq \mathbb{E}|x_t^2 - \phi^{2G}x_{t-G}^2|\\
&\leq \mathbb{E}D_t^2 + 2\phi^{2G}\mathbb{E}|x_{t-G}D_t|\\
&\leq GM_1 + 2M_1 M_2^{2/(C_0+2)}\sqrt{G(t-G)}.
\end{aligned} \tag{96}$$

Step 2: bound $\mathbb{E}|\mathbb{E}(x_{t-G}^2 \mid z_{1t}) - \mathbb{E}x_{t-G}^2|$.

Let $r = 1 + C_0/2$. We apply Lemma 1 with $X = x_{t-G}^2 - \mathbb{E}x_{t-G}^2$, $Y = 1$, $\mathcal{F} = \sigma(\{u_1, ..., u_{t-G}\})$, $\mathcal{G} = \sigma(z_{1t})$, $p = 1$ and $q = \infty$. It follows that

$$\mathbb{E}\left|\mathbb{E}\left(x_{t-G}^2 - \mathbb{E}x_{t-G}^2 \mid z_{1t}\right)\right| \leq 8[\alpha(G)]^{1-1/r}\|x_{t-G}^2 - \mathbb{E}x_{t-G}^2\|_{L^r(\mathbb{P})}.$$

Notice that

$$\|x_{t-G}^2 - \mathbb{E}x_{t-G}^2\|_{L^r(\mathbb{P})} \leq 2\|x_{t-G}^2\|_{L^r(\mathbb{P})} = 2\left(\mathbb{E}|x_{t-G}|^{2+C_0}\right)^{1/r} \overset{(i)}{\leq} 2M_2^{1/r}M_1(t-G),$$

where (i) holds by (94). The above two displays imply that

$$\mathbb{E}\left|\mathbb{E}\left(x_{t-G}^2 - \mathbb{E}x_{t-G}^2 \mid z_{1t}\right)\right| \leq 16[\alpha(G)]^{C_0/(C_0+2)} M_2^{2/(C_0+2)} M_1(t-G). \tag{97}$$

Now we observe that

$$\begin{aligned}
&\mathbb{E}\left|\mathbb{E}(x_t^2 \mid z_{1t}) - \mathbb{E}x_t^2\right| \\
&\leq \mathbb{E}\left|\mathbb{E}(x_t^2 \mid z_{1t}) - \mathbb{E}(x_{t-G}^2 \mid z_{1t})\right| + \mathbb{E}\left|\mathbb{E}(x_{t-G}^2 \mid z_{1t}) - \mathbb{E}x_{t-G}^2\right| \mathbb{E}\left|\mathbb{E}x_{t-G}^2 - \mathbb{E}(x_{t-G}^2 \mid z_{1t})\right| \\
&\overset{(i)}{\leq} 2GM_1 + 4M_1 M_2^{2/(C_0+2)} \sqrt{G(t-G)} + 16[\alpha(G)]^{C_0/(C_0+2)} M_2^{2/(C_0+2)} M_1(t-G),
\end{aligned}$$

where (i) holds by (96) and (97). The desired result follows. □

Lemma 19. *Let the assumptions of Proposition 8 hold. Then there exists a constant $K_4 > 0$ such that*

$$\mathbb{E}\left|\mathbb{E}\left(x_t \varepsilon_{t+1} \mid z_{1t}\right)\right| \leq K_4 \sum_{i=0}^{t-1} [\alpha(i)]^{(C_0+1)/(C_0+2)}.$$

Proof. Let $R = \mathbb{E}(u_{t-i} \mid z_{1t}, \varepsilon_{t+1})$. Then

$$\begin{aligned}
\mathbb{E}\left|\mathbb{E}\left(u_{t-i}\varepsilon_{t+1} \mid z_{1t}\right)\right| = \mathbb{E}\left|\mathbb{E}(R \mid z_{1t})\right| &\leq \mathbb{E}\left(\mathbb{E}\left(|R| \mid z_{1t}\right)\right) \\
&= \mathbb{E}\left|\mathbb{E}(u_{t-i} \mid z_{1t}, \varepsilon_{t+1})\right| \\
&\overset{(i)}{\leq} 8\left[\alpha(i)\right]^{(C_0+1)/(C_0+2)} C_2^{1/(C_0+2)},
\end{aligned}$$

where (i) follows by applying Lemma 1 (with $X = u_{t-i}$, $Y = 1$, $\mathcal{F} = \sigma(u_{t-i})$, $\mathcal{G} = \sigma(\varepsilon_{t+1}, z_{1t})$, $p = 2 + C_0$ and $q = \infty$). Therefore,

$$\begin{aligned}
\mathbb{E}\left|\mathbb{E}\left(x_t\varepsilon_{t+1} \mid z_{1t}\right)\right| = \mathbb{E}\left|\sum_{i=0}^{t-1} \phi^i \mathbb{E}\left(u_{t-i}\varepsilon_{t+1} \mid z_{1t}\right)\right| &\overset{(i)}{\leq} \sum_{i=0}^{t-1} \mathbb{E}\left|\mathbb{E}\left(u_{t-i}\varepsilon_{t+1} \mid z_{1t}\right)\right| \\
&\leq 8C_2^{1/(C_0+2)} \sum_{i=0}^{t-1} [\alpha(i)]^{(C_0+1)/(C_0+2)},
\end{aligned}$$

where (i) follows by $|\phi| \leq 1$. The proof is complete. □

Proof of Proposition 8 . Since $\mathbb{E}x_t\varepsilon_{t+1} = 0$, we have $\mathbb{E}\Delta L_{t+1} = \beta^2 \mathbb{E}x_t^2$. Notice that $x_t = \sum_{i=0}^{t-1} \phi^i u_{t-i}$. Therefore, $\mathbb{E}x_t^2 = \sum_{i=0}^{t-1} \phi^{2i} \mathbb{E}u_{t-i}^2 \in [C_1 \sum_{i=0}^{t-1} \phi^{2i},\ C_2 \sum_{i=0}^{t-1} \phi^{2i}]$.

Since $\sum_{i=0}^{t-1}\phi^{2i} = (1-\phi^{2t})/(1-\phi^2)$ and $\phi = \exp(-c_\phi/T)$, we have that $t^{-1}\sum_{i=0}^{t-1}\phi^{2i} \to 1/2$ as $t, T \to \infty$. It follows by $n \le t \le T$ and $n \asymp T$ that $\mathbb{E}x_t^2 \asymp T$. The first part is proved.

For the second part, notice that

$$\begin{aligned}
&\mathbb{E}\left|\mathbb{E}(\Delta L_{t+1} \mid z_{1t}) - \mathbb{E}(\Delta L_{t+1})\right| \\
&= \mathbb{E}\left|\beta^2\left[\mathbb{E}(x_t^2 \mid z_{1t}) - \mathbb{E}x_t^2\right] + 2\beta\mathbb{E}(x_t\varepsilon_{t+1} \mid z_{1t})\right| \\
&\le \beta^2\mathbb{E}\left|\mathbb{E}(x_t^2 \mid z_{1t}) - \mathbb{E}x_t^2\right| + 2\beta\mathbb{E}\left|\mathbb{E}(x_t\varepsilon_{t+1} \mid z_{1t})\right| \\
&\overset{(i)}{\le} \beta^2 \min_{1\le G\le t}\left(K_1 G + K_2\sqrt{G(t-G)} + K_3[\alpha(G)]^{C_0/(C_0+2)}(t-G)\right) \\
&\quad + K_4\beta\sum_{i=0}^{t-1}[\alpha(i)]^{(C_0+1)/(C_0+2)}
\end{aligned} \tag{98}$$

for some constants $K_1, ..., K_4 > 0$, where (i) follows by Lemmas 18 and 19.

Let $W_t = \mathbb{E}(\Delta L_{t+1} \mid z_{1t})$. Thus, $\mathbb{E}W_t = \mathbb{E}\Delta L_{t+1} = \beta^2\mathbb{E}x_t^2 > 0$. By Markov's inequality, we have

$$\begin{aligned}
\mathbb{P}(W_t \le 0) &= \mathbb{P}(W_t - \mathbb{E}W_t \le -\mathbb{E}W_t) \\
&\le \mathbb{P}(|W_t - \mathbb{E}W_t| \ge \mathbb{E}W_t) \\
&\le \frac{\mathbb{E}|W_t - \mathbb{E}W_t|}{\mathbb{E}W_t} \\
&\overset{(i)}{\le} T^{-1}\min_{1\le G\le t}\left(K_1 G + K_2\sqrt{G(t-G)} + K_3[\alpha(G)]^{C_0/(C_0+2)}(t-G)\right) \\
&\quad + c_\beta^{-1}T^{\alpha-1}K_3\sum_{i=0}^{t-1}[\alpha(i)]^{(C_0+1)/(C_0+2)}
\end{aligned}$$

where (i) follows by (98) and $\mathbb{E}W_t = \beta^2\mathbb{E}x_t^2$ with $\mathbb{E}x_t^2 \asymp T$. This proves the second claim.

To see the third claim, let $A = \mathbb{E}W_t = \mathbb{E}\Delta L_{t+1} = \beta^2\mathbb{E}x_t^2$. Then

$$\begin{aligned}
\mathbb{E}|W_t\mathbf{1}\{W_t > 0\} - a| &= \mathbb{E}|(W_t - a)\mathbf{1}\{W_t > 0\} - a\mathbf{1}\{W_t \le 0\}| \\
&\le \mathbb{E}|W_t - a| + a\mathbb{P}(W_t \le 0) \\
&\overset{(i)}{\le} 2\mathbb{E}|W_t - a|,
\end{aligned}$$

where (i) holds by $\mathbb{P}(W_t \leq 0) \leq \mathbb{E}|W_t - a|/a$ from the proof of the second claim. By (98), $\beta \asymp T^{-\alpha}$, $\mathbb{E}x_t^2 \asymp T$ and $a \asymp T^{1-2\alpha}$, the third claim follows. The proof is complete. □

References

Andrews, D. W. (1988). Laws of large numbers for dependent non-identically distributed random variables. *Econometric theory*, 4(03):458–467.

Athreya, K. B. and Lahiri, S. N. (2006). *Measure Theory and Probability Theory*. Springer Science & Business Media.

Bonhomme, S. and Manresa, E. (2015). Grouped patterns of heterogeneity in panel data. *Econometrica*, 83(3):1147–1184.

Bradley, R. C. (2007). *Introduction to strong mixing conditions*, volume 1. Kendrick Press Heber City.

Capistrán, C. (2008). Bias in Federal Reserve inflation forecasts: Is the Federal Reserve irrational or just cautious? *Journal of Monetary Economics*, 55(8):1415–1427.

Chen, X., Shao, Q.-M., and Wu, W. B. (2016). Self-normalized cramér-type moderate deviations under dependence. *The Annals of Statistics*, 44(4):1593–1617.

Chong, Y. Y. and Hendry, D. F. (1986). Econometric evaluation of linear macroeconomic models. *The Review of Economic Studies*, 53(4):671–690.

Clark, T. and McCracken, M. (2013). Advances in forecast evaluation. In Elliott, G. and Timmermann, A., editors, *Handbook of Economic Forecasting*, volume 2, part B, pages 1107–1201. Elsevier.

Clark, T. E. and McCracken, M. W. (2001). Tests of equal forecast accuracy and encompassing for nested models. *Journal of econometrics*, 105(1):85–110.

Diebold, F. X. (2015). Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of diebold-mariano tests. *Journal of Business & Economic Statistics*, 33(1):1–8.

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

Fan, J., Liao, Y., and Mincheva, M. (2011). High dimensional covariance matrix estimation in approximate factor models. *Annals of statistics*, 39(6):3320.

Giacomini, R. and Rossi, B. (2009). Detecting and predicting forecast breakdowns. *Review of Economic Studies*, 76(2):669–705.

Giacomini, R. and White, H. (2006). Tests of conditional predictive ability. *Econometrica*, 74(6):1545–1578.

Granger, C. W. J. and Newbold, P. (1977). *Forecasting economic time series*. Academic Press.

Gut, A. (2013). *Probability A Graduate Course*. Springer.

Hansen, P. R. and Timmermann, A. (2015a). Comment on comparing predictive accuracy, twenty years later. *Journal of Business and Economic Statistics*, 33:17–21.

Hansen, P. R. and Timmermann, A. (2015b). Equivalence between out-of-sample forecast comparisons and wald statistics. *Econometrica*, 83(6):2485–2505.

Hirano, K. and Wright, J. H. (2017). Forecasting with model uncertainty: Representations and risk reduction. *Econometrica*, 85(2):617–643.

Jurado, K., Ludvigson, S. C., and Ng, S. (2015). Measuring uncertainty. *The American Economic Review*, 105(3):1177–1216.

McCracken, M. W. (2007). Asymptotics for out of sample tests of granger causality. *Journal of Econometrics*, 140(2):719–752.

Peña, V. H., Lai, T. L., and Shao, Q.-M. (2008). *Self-normalized processes: Limit theory and Statistical Applications*. Springer Science & Business Media.

Pesaran, M. H., Pettenuzzo, D., and Timmermann, A. (2006). Forecasting time series subject to multiple structural breaks. *Review of Economic Studies*, 73(4):1057–1084.

Pettenuzzo, D. and Timmermann, A. (2015). Forecasting macroeconomic variables under model instability. *Journal of Business and Economic Statistics*. Available at SSRN 2603879.

Rapach, D. E., Strauss, J. K., and Zhou, G. (2010). Out-of-sample equity premium prediction: Combination forecasts and links to the real economy. *Review of Financial Studies*, 23(2):821–862.

Romer, C. D. and Romer, D. H. (2000). Federal reserve information and the behavior of interest rates (digest summary). *American Economic Review*, 90(3):429–457.

Rossi, B. (2013). Advances in forecasting under instability. In Elliott, G. and Timmermann, A., editors, *Handbook of Economic Forecasting*, volume 2, part B, chapter 21, pages 1203–1324. Elsevier.

Rossi, B. and Sekhposyan, T. (2013). Conditional predictive density evaluation in the presence of instabilities. *Journal of Econometrics*, 177(2):199–212.

Rossi, B. and Sekhposyan, T. (2015). Forecast rationality tests in the presence of instabilities, with applications to Federal Reserve and survey forecasts. *Journal of Applied Econometrics*.

Stock, J. H. and Watson, M. W. (1996). Evidence on structural instability in macroeconomic time series relations. *Journal of Business & Economic Statistics*, 14(1):11–30.

Stock, J. H. and Watson, M. W. (2003). Forecasting output and inflation: The role of asset prices. *Journal of Economic Literature*, 41(3):788–829.

Stock, J. H. and Watson, M. W. (2007). Why has us inflation become harder to forecast? *Journal of Money, Credit and banking*, 39(s1):3–33.

Stock, J. H. and Watson, M. W. (2010). Modeling inflation after the crisis. Technical report, National Bureau of Economic Research.

Timmermann, A. (2007). An evaluation of the world economic outlook forecasts. *IMF Staff Papers*, 54(1):1–33.

Timmermann, A. and Zhu, Y. (2016). Tests of forecasting performance and choice of estimation window. *UCSD Working Paper*.

Welch, I. and Goyal, A. (2008). A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies*, 21(4):1455–1508.

West, K. D. (1996). Asymptotic inference about predictive ability. *Econometrica: Journal of the Econometric Society*, pages 1067–1084.

Table 1: Predictive performance of nested models and the switching approach

$\alpha_z \backslash \alpha_x$	0	0.1	0.25	0.4	0.5	0.75	1
	A: $Pr(J_T^{Small-Big} > 1.64)$						
	1.000	1.000	0.995	0.463	0.110	0.004	0.001
	B: $Pr(J_T^{Small-SW} > 1.64)$						
0.0	1.000	1.000	1.000	0.999	0.965	0.623	0.519
0.1	1.000	1.000	1.000	0.973	0.851	0.369	0.269
0.25	1.000	1.000	0.996	0.677	0.358	0.094	0.049
0.5	1.000	1.000	0.993	0.419	0.121	0.014	0.010
0.75	1.000	1.000	0.991	0.393	0.087	0.008	0.005
1.0	1.000	1.000	0.994	0.389	0.092	0.009	0.005
	C: $Pr(J_T^{Big-SW} > 1.64)$						
0.0	0.297	0.729	0.990	0.998	0.977	0.820	0.762
0.1	0.058	0.118	0.360	0.693	0.753	0.630	0.597
0.25	0.005	0.018	0.036	0.097	0.166	0.284	0.280
0.5	0.000	0.004	0.007	0.010	0.026	0.141	0.170
0.75	0.000	0.002	0.004	0.006	0.018	0.126	0.167
1.0	0.000	0.003	0.005	0.004	0.016	0.122	0.154

This table reports result from 5,000 Monte Carlo simulations using a sample size of $(n, m, p) = (100, 100, 200)$. The table reports the probability of rejecting the null of equal MSE loss against a one-sided alternative that the MSE of one set of forecasts exceeds the MSE from a second set of forecasts.
Data are generated from the linear regression model

$$y_{t+1} = \beta x_t + \varepsilon_{t+1},$$

where the predictor $x_t \sim U(-1, 1)$ is iid. Let $s_{t+1} \in \{0, 1\}$ be a binary indicator such that $P(s_{t+1} = 1 \mid x_t > 0) = \mu + \delta$ and $P(s_{t+1} = 1 \mid x_t \leq 0) = \mu - \delta$, where $\mu = 0.5$. The residuals ε_{t+1} are generated as

$$\varepsilon_{t+1} = s_{t+1} Q_{1,t} + (1 - s_{t+1}) Q_{2,t},$$

where $Q_{1,t}$ and $Q_{2,t}$ are $N(0, 1)$ independent of each other and of s_{t+1} and x_t. The monitoring instrument is generated as

$$z_{1t} = a_1 Q_{1,t} + a_2 Q_{2,t},$$

where $a_1 = 1$ and $a_2 = -1$. It is easy to see that $Ex_t\varepsilon_{t+1} = Ez_{1t}\varepsilon_{t+1} = E\varepsilon_{t+1} = Ez_{1t} = Ex_t = 0$ and

$$Corr(x_t\varepsilon_{t+1}, z_{1t}) = \sqrt{\frac{3}{2}}\delta.$$

We choose $\beta = 3n^{-\alpha_x}$ and δ to satisfy $Corr(x_t\varepsilon_{t+1}, z_{1t}) = 0.6n^{-\alpha_z}$. The big forecasting model always includes x_t as a predictor whereas the small model predicts zero. The switching rule regresses the squared error differences of the small and big model on an intercept and the monitoring instrument, z_{1t}, and chooses the model that is expected to generate the smallest expected loss for the next period.

Table 2: Predictive performance of the switching approach compared to model combination, pretesting and a model augmented with the monitoring instrument

$\alpha_z \backslash \alpha_x$	0	0.1	0.25	0.4	0.5	0.75	1
A: $Pr(J_T^{Comb-SW} > 1.64)$							
0.0	1.000	1.000	1.000	1.000	0.977	0.746	0.685
0.1	1.000	0.994	0.914	0.903	0.850	0.552	0.475
0.25	1.000	0.993	0.507	0.263	0.226	0.167	0.151
0.5	1.000	0.996	0.366	0.042	0.025	0.051	0.061
0.75	1.000	0.995	0.365	0.025	0.015	0.039	0.050
1.0	1.000	0.996	0.369	0.025	0.010	0.042	0.058
B: $Pr(J_T^{Pretest-SW} > 1.64)$							
0.0	0.300	0.734	0.988	1.000	0.969	0.653	0.558
0.1	0.057	0.118	0.365	0.798	0.859	0.452	0.347
0.25	0.005	0.021	0.033	0.272	0.391	0.150	0.101
0.5	0.001	0.004	0.005	0.094	0.136	0.045	0.029
0.75	0.000	0.003	0.005	0.076	0.111	0.037	0.024
1.0	0.000	0.003	0.005	0.063	0.095	0.028	0.023
C: $Pr(J_T^{Augment-SW} > 1.64)$							
0.0	0.369	0.768	0.967	0.945	0.867	0.616	0.565
0.1	0.160	0.166	0.441	0.642	0.641	0.478	0.456
0.25	0.169	0.112	0.086	0.183	0.241	0.314	0.303
0.5	0.182	0.146	0.057	0.060	0.102	0.220	0.235
0.75	0.193	0.138	0.057	0.050	0.088	0.209	0.236
1.0	0.180	0.151	0.053	0.044	0.085	0.206	0.233

Using the nested model setup from Table 1, this table reports the probability of rejecting the null of equal MSE loss against a one-sided alternative that the MSE of one set of forecasts exceeds the MSE from a second set of forecasts. In each panel, we compare the MSE performance of the of the switching approach to that of an equal-weighted combination (Panel 1), an approach that includes a predictor in the forecasting model if its regression coefficient is statistically significant (pretest, in Panel 2) and forecasts from an augmented model that includes both the predictor, x_t, and the monitoring instrument, z_{1t}, in the forecasting model. In each case, large values of the rejection probability indicates that the switching approach performs better (produces a smaller MSE) than the alternative approach.

Table 3: Pairwise comparisons of predictive performance for the non-nested case (with β_2's sign flipped)

$Pr(J_T^{j_1-j_2} > 1.64)$

	$(j_1,j_2)=(1,2)$				$(j_1,j_2)=(1,SW)$				$(j_1,j_2)=(2,SW)$			
A: $(\alpha_{z,1},\alpha_{z,2})=(0,0)$												
$\alpha_{x,1}\backslash\alpha_{x,2}$	0	0.25	0.5	1	0	0.25	0.5	1	0	0.25	0.5	1
0.0	0.063	0.000	0.000	0.000	0.716	0.004	0.002	0.001	0.700	1.000	1.000	1.000
0.25	1.000	0.054	0.000	0.000	1.000	0.965	0.327	0.166	0.006	0.961	1.000	0.999
0.5	1.000	0.985	0.031	0.004	1.000	1.000	0.804	0.371	0.003	0.332	0.812	0.527
1.0	1.000	0.996	0.083	0.012	1.000	1.000	0.529	0.256	0.001	0.166	0.367	0.249
B: $(\alpha_{z,1},\alpha_{z,2})=(0,1)$												
$\alpha_{x,1}\backslash\alpha_{x,2}$	0	0.25	0.5	1	0	0.25	0.5	1	0	0.25	0.5	1
0.0	0.059	0.000	0.000	0.000	0.202	0.002	0.001	0.002	0.204	1.000	1.000	1.000
0.25	1.000	0.054	0.000	0.000	1.000	0.379	0.147	0.151	0.000	0.378	0.998	1.000
0.5	1.000	0.987	0.032	0.004	1.000	0.983	0.256	0.268	0.000	0.008	0.299	0.465
1.0	1.000	0.996	0.081	0.012	1.000	0.993	0.143	0.107	0.000	0.005	0.068	0.133
C:$(\alpha_{z,1},\alpha_{z,2})=(0.5,0.5)$												
$\alpha_{x,1}\backslash\alpha_{x,2}$	0	0.25	0.5	1	0	0.25	0.5	1	0	0.25	0.5	1
0.0	0.058	0.000	0.000	0.000	0.050	0.000	0.000	0.000	0.050	1.000	1.000	1.000
0.25	1.000	0.067	0.000	0.000	1.000	0.058	0.005	0.004	0.000	0.049	0.977	0.994
0.5	1.000	0.987	0.033	0.004	1.000	0.973	0.037	0.008	0.000	0.005	0.029	0.068
1.0	1.000	0.998	0.076	0.013	1.000	0.994	0.065	0.017	0.000	0.003	0.010	0.016
D: $(\alpha_{z,1},\alpha_{z,2})=(1,1)$												
$\alpha_{x,1}\backslash\alpha_{x,2}$	0	0.25	0.5	1	0	0.25	0.5	1	0	0.25	0.5	1
0.0	0.059	0.000	0.000	0.000	0.053	0.000	0.000	0.000	0.050	1.000	1.000	1.000
0.25	1.000	0.054	0.000	0.000	1.000	0.044	0.004	0.003	0.000	0.050	0.979	0.992
0.5	1.000	0.984	0.032	0.004	1.000	0.974	0.031	0.008	0.000	0.004	0.033	0.062
1.0	1.000	0.997	0.079	0.014	1.000	0.994	0.063	0.015	0.000	0.003	0.008	0.019

This table reports rejection probabilities of the null of equal MSE performance of models j_1 and j_2. Data are generated according to the non-nested model

$$y_{t+1} = \beta_1 x_{1,t} + \beta_2 x_{2,t} + \varepsilon_{t+1},$$

where x_{1t} and x_{2t} are a set of predictor variables that are known at time t. Model 1 takes the form $y_{t+1} = \beta_1 x_{1,t} + \varepsilon_{1t+1}$, while model 2 takes the form $y_{t+1} = \beta_2 x_{2,t} + \varepsilon_{2t+1}$. The strength of the predictors in models 1 and 2 is parameterized as $\beta_j = c_{\beta,j} n^{-\alpha_{x,j}}$, while the accuracy of the monitoring instrument is captured as $corr(x_{j,t}\varepsilon_{t+1}, z_{1t}) = c_{\rho,j} m^{-\alpha_{z,j}}$. We report the outcome of a one-sided test of the null of equal predictive performance against the alternative that $Pr(J_T^{j_1-j_2} > 1.64)$. Higher values of the probabilities indicate that model $j_2$produces a lower MSE than model j_1. All results are based on 5,000 MC simulations and use a sample size of $(n, m, p) = (100, 100, 200)$.

Table 4: Diebold-Mariano tests for equal forecasting performance: Greenbook versus SPF forecasts

$h \setminus Z$	Uncond	UG		U		$\overline{\Delta L}_{t-3:t}$		$\overline{\triangle(\hat{y}^2_{1,s-3:t} - \hat{y}^2_{2,s-3:t})}$	
	ΔL^{GB-SPF}	ΔL^{GB-SW}	ΔL^{SPF-SW}	ΔL^{GB-SW}	ΔL^{SPF-SW}	ΔL^{GB-SW}	ΔL^{SPF-SW}	ΔL^{GB-SW}	ΔL^{SPF-SW}
				A: Final, revised data ($m = 40$)					
1	-1.97	1.26	1.80	1.21	1.68	1.05	2.19	1.50	2.07
2	-2.19	0.62	2.16	1.54	2.27	1.53	2.94	1.29	2.02
3	-2.25	1.56	2.23	0.97	1.94	1.74	2.36	1.49	1.24
4	-2.26	1.62	2.06	1.91	2.34	1.90	2.86	1.78	1.02
				B: Final, revised data ($m = 60$)					
1	-1.97	1.43	1.35	1.30	1.30	1.80	1.51	1.50	0.97
2	-2.19	1.30	1.53	1.47	1.29	1.51	2.01	1.30	0.92
3	-2.25	1.56	0.80	1.80	1.21	1.57	1.34	1.69	-0.13
4	-2.26	1.63	0.41	2.23	0.73	1.79	2.04	1.46	0.29
				C: Vintage data ($m = 40$)					
1	-3.24	1.72	3.42	-1.10	2.27	-0.62	2.43	-0.56	2.68
2	-2.29	-0.02	2.14	-0.19	1.81	-0.04	2.06	0.42	1.94
3	-2.41	0.36	2.83	0.92	2.42	1.69	2.92	1.43	2.33
4	-2.56	0.60	2.36	0.54	2.18	1.43	3.14	0.89	1.91
				D: Vintage data ($m = 60$)					
1	-3.24	1.38	2.44	-1.07	2.08	-1.51	1.75	-0.83	1.87
2	-2.29	-0.06	1.81	1.12	2.05	1.15	2.46	NA	1.54
3	-2.41	0.79	0.97	1.27	1.13	1.66	2.09	0.76	0.97
4	-2.56	0.03	1.22	-0.06	1.19	1.67	2.70	-0.31	1.20

This table reports Diebold-Mariano t-tests for equal mean squared error performance for the Greenbook and mean SPF forecasts of the GDP deflator as well as for various rules that switch between these forecasts. All forecasts are reported quarterly using forecast horizons (h, listed in the rows) ranging from one through four quarters. The first column reports the Diebold-Mariano t-test for the unconditional forecasting performance with negative values suggesting that the Greenbook forecasts are more accurate than the SPF forecasts. Subsequent columns compare the MSE performance of various switching rules against the Greenbook forecasts (labeled ΔL^{GB-SW}) or against the SPF forecasts (ΔL^{SPF-SW}) with positive values indicating that the switching rule performs best. The switching rule uses a rolling window of $m = 40$ quarterly observations (panel A) or $m = 60$ observations (panel B) to regress loss differences on an intercept and the test variable, Z, which is either the unemployment gap (UG), the macroeconomic uncertainty measure of Jurado et al. (2015) (U), the lagged loss differential averaged over the most recent four quarters ($\overline{\Delta L}$), or the difference between the squared forecasts, averaged over the most recent four quarters, $\overline{\triangle(\hat{y}^2_{1,s} - \hat{y}^2_{2,s})}$. Results in panels A and B use the most recent vintage of the GDP deflator to measure the "actual" value while results in panels C and D use real-time vintages. The sample period is 1968Q4-2010Q4.

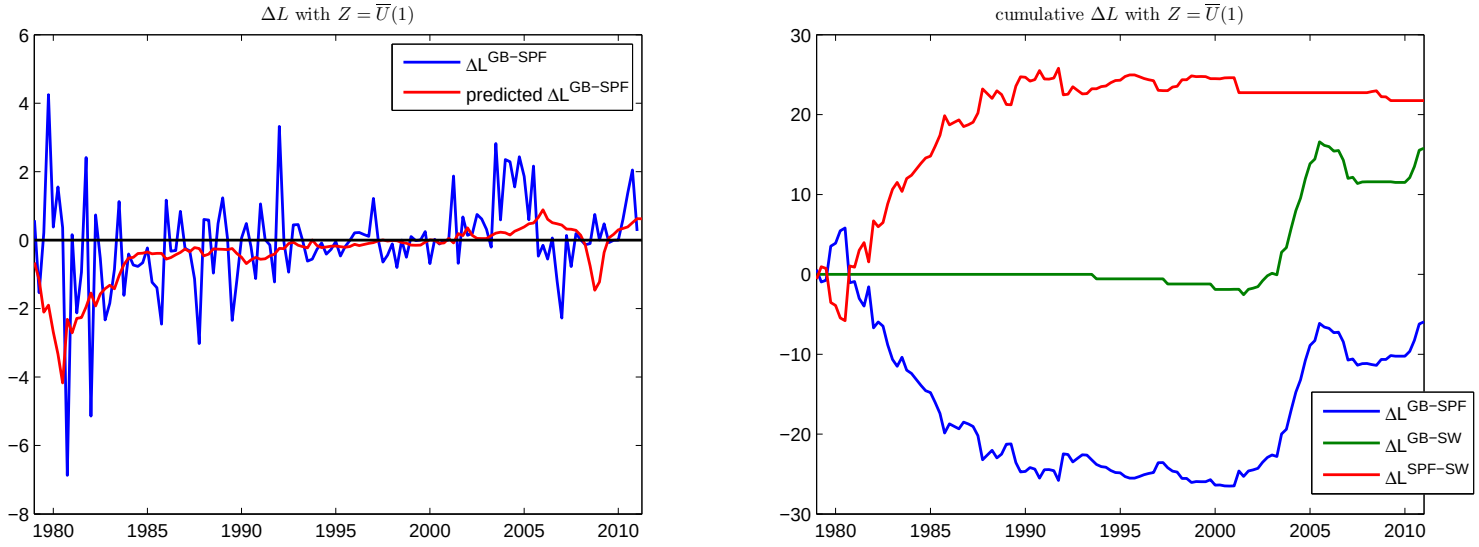
Table 5: Empirical results from forecasts of stock returns

A:Model combination

Z	t_0	t_1	R^2	GW
1	1.82*			
$\overline{\Delta L}$	2.70***	-0.68	0.44	0.01**
UG	2.87***	-1.31	0.38	0.01**
$\hat{y}_1^2 - \hat{y}_2^2$	0.32	-2.31***	1.61	0.01**
$U(1)$	-0.30	0.39	0.11	0.13
	B: T-bill rate			
1	-1.11			
$\overline{\Delta L}$	-1.01	-0.42	0.16	0.54
UG	0.59	-2.63***	0.47	0.04**
$\hat{y}_1^2 - \hat{y}_2^2$	-0.37	-1.96**	1.39	0.24
$U(1)$	0.08	-0.10	0.00	0.98

This table reports the outcome of conditional tests of equal forecasting performance that compare forecasts of monthly excess returns on a U.S. stock market portfolio from a (small, S) prevailing mean model ($y_{t+1} = \beta_0 + \varepsilon_{t+1}$) to forecasts from a (big, B) model with time-varying predictors ($y_{t+1} = \beta_0 + \beta_1 x_t + \varepsilon_{t+1}$). In both cases, the parameters of the forecasting model are estimated using a rolling window with $n = 240$ (20 years) of observations, generating a sequence of loss differentials $\Delta L_{t+1} = e^2_{S,t+1} - e^2_{B,t+1}$. We show tests of the null $E[\Delta L_{t+1}|Z_t] = 0$ versus a two-sided alternative under squared error loss, using instruments $Z_t = 1$ (corresponding to a test of equal unconditional expected loss), $Z_t = \overline{\Delta L}_t$ (the average loss differential over the preceding 12 months), $Z_t = \overline{\hat{y}^2_{S,s} - \hat{y}^2_{B,s}}$, the lagged difference in squared forecasts, $Z_t = ug_t$, the unemployment gap variable of Stock and Watson (2010), or $Z_t = U(1)_t$, the lagged one-month uncertainty measure of Jurado, Ludvigson and Ng (2016). Numbers shown are t-statistics on θ_0 and θ_1 from regressions $\Delta L_{t+1} = \theta_0 + \theta_1 z_{1t} + \epsilon_{t+1}$, the R^2 from this regression and the Giacomini-White (GW) test of conditional predictability. In all cases the small (benchmark) forecasting model is a prevailing mean model while the big model is a forecast combination (Panel A) or a univariate forecasting model that includes a T-bill rate (Panel B).

Figure 1: Mean squared error differences and forecasting performance for Greenbook, SPF and switching rule forecasts



This figure shows the the squared error differences for Greenbook versus SPF forecasts (blue line, left panel) along with conditional forecasts of these loss differences using the unemployment gap (red line, left panel) as a monitoring instrument. The right panel shows the cumulative sums of squared forecast errors from comparisons of Greenbook and SPF forecasts (blue line), switching rule versus Greenbook forecasts (green line) and switching rule versus SPF forecasts (red line).

Figure 2: Split of sample (T) into estimation (n), monitoring (m), and evaluation ($n+m+p$) parts

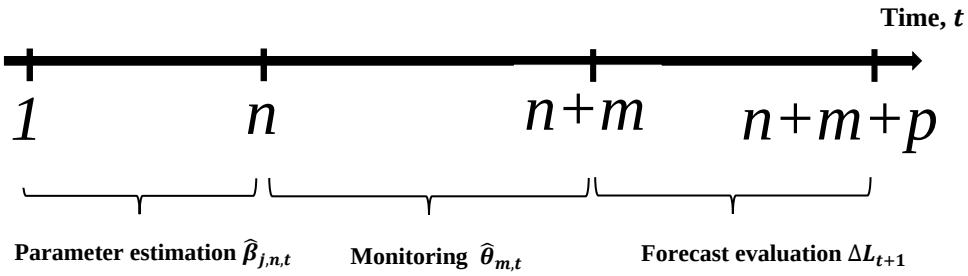
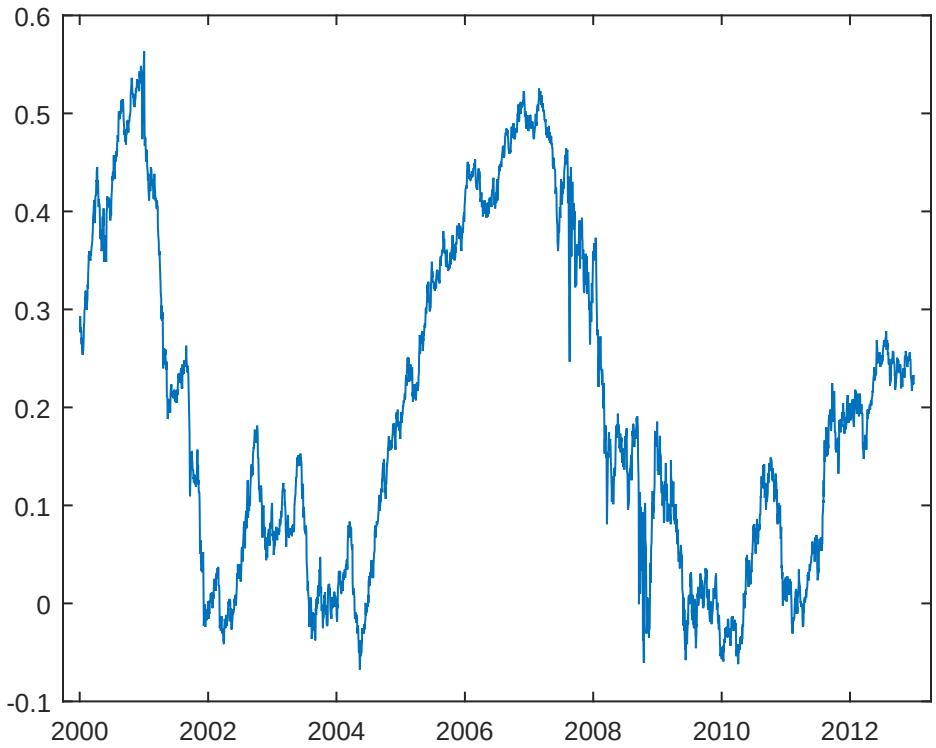
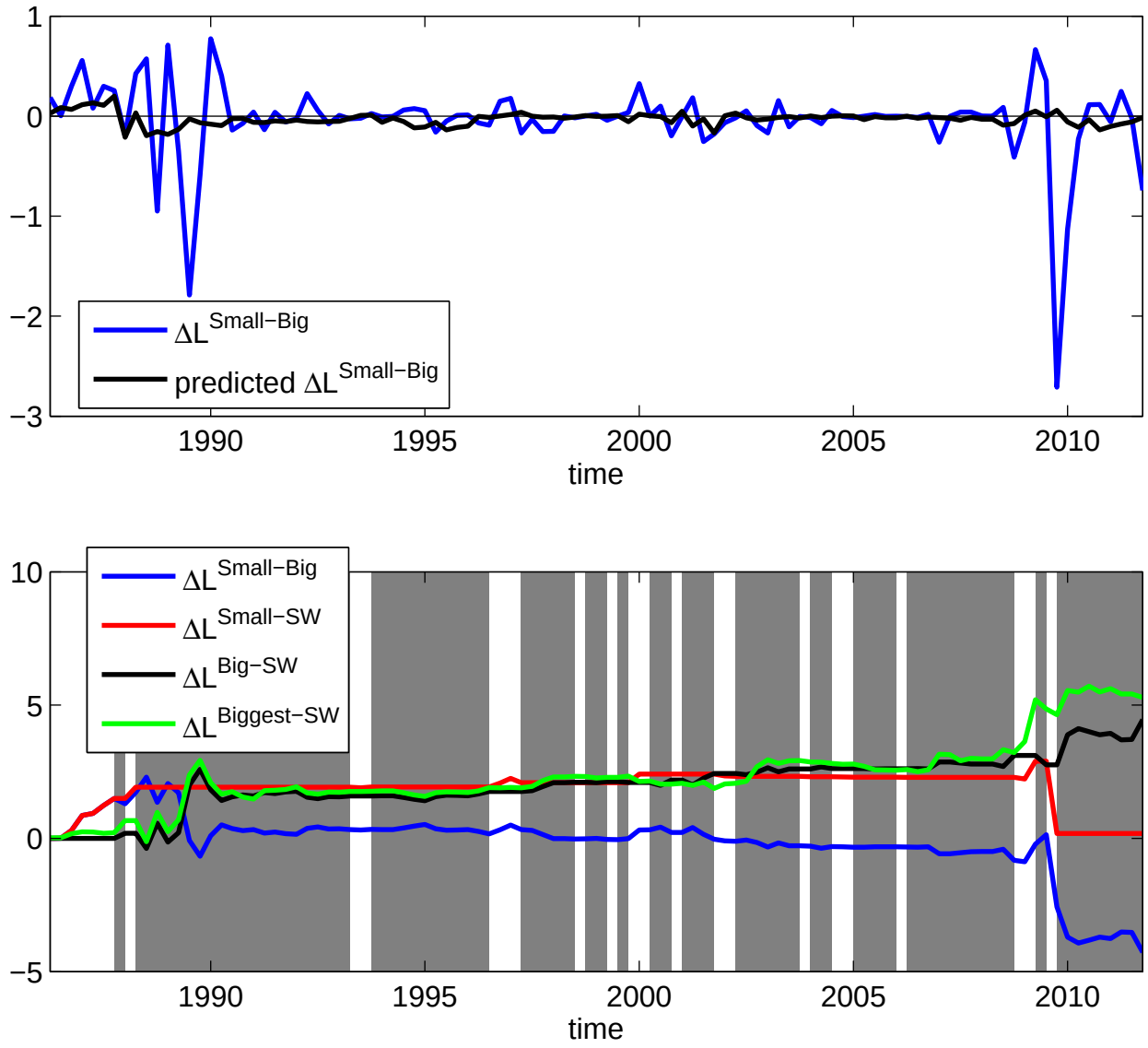


Figure 3: Daily forecasts of squared error difference for Greenbook versus SPF inflation forecasts



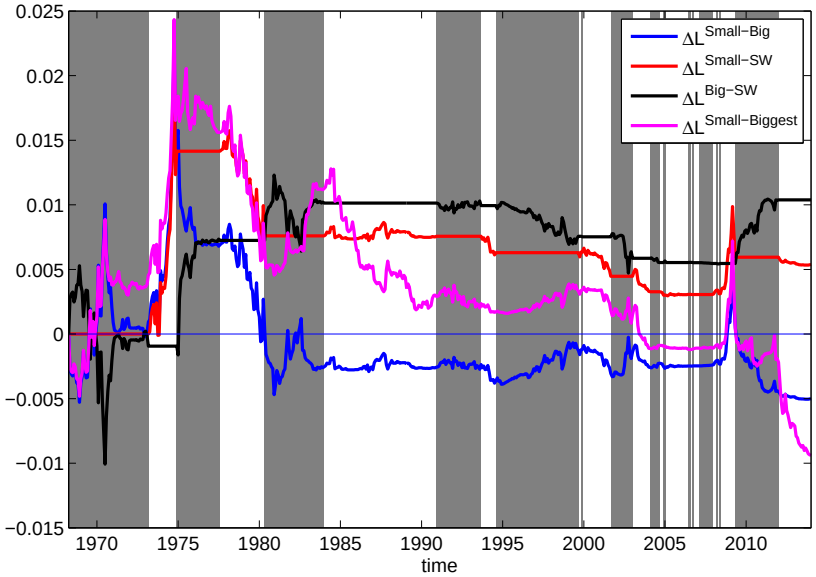
This figure plots daily estimates of the squared error difference between Greenbook and SPF inflation forecasts, computed using daily values of the term spread. Positive values indicate that the Greenbook forecasts are expected to be less accurate than the SPF forecasts, while negative estimates suggest the reverse.

Figure 4: Performance of Nested Forecasting Models for US Inflation



Forecasting performance for models of US inflation. The top panel plots the conditionally expected squared error differential for an AR(4) model (small) versus a model that adds a principal component extracted from a large cross-section of macroeconomic variables to the AR(4) model (big). The conditional expectation is computed using the unemployment gap as a tester. The bottom panel plots the cumulative sum of squared error differences for the big model versus the small model (blue line), the switching rule versus the small model (red), the switching rule versus the big model (black) and the switching rule versus an augmented model that adds the tester (unemployment gap) to the big forecasting model.

Figure 5: Cumulative Sum of Squared Error Differences for Different Approaches to Forecasting US Stock Returns



This figure plots the cumulative sums of squared forecast error differentials for big versus small forecasting model (blue line), switching rule versus small forecasting model (red), switching rule versus big model (black) and big forecasting model augmented with the monitoring instrument versus small model (purple). Positive and increasing values of the lines suggest that the first model produces smaller squared forecast errors than the second model, while negative and decreasing values suggest the opposite. Gray areas show when the big forecasting model is expected to generate the smallest squared error loss. The big forecasting model includes a constant and the lagged T-bill rate, while the small forecasting model only includes a constant.