working paper 21 06

All Forecasters Are Not the Same: Time-Varying Predictive Ability across Forecast Environments

Robert Rich and Joseph Tracy

FEDERAL RESERVE BANK OF CLEVELAND

ISSN: 2573-7953

Working Paper 21-06 February 2021

All Forecasters Are Not the Same: Time-Varying Predictive Ability across Forecast Environments

Robert Rich and Joseph Tracy

This paper examines data from the European Central Bank's Survey of Professional Forecasters to investigate whether participants display equal predictive performance. We use panel data models to evaluate point- and density-based forecasts of real GDP growth, inflation, and unemployment. The results document systematic differences in participants' forecast accuracy that are not time invariant, but instead vary with the difficulty of the forecasting environment. Specifically, we find that some participants display higher relative accuracy in tranquil environments, while others display higher relative accuracy in volatile environments. We also find that predictive performance is positively correlated across target variables and horizons, with density forecasts generating stronger correlation patterns. Taken together, the results support the development of expectations models featuring persistent heterogeneity.

JEL codes:C12; C33; C53.

Keywords: professional forecasters, survey data, forecast accuracy, point forecasts, density forecasts, persistent heterogeneity.

Suggested citation: Rich, Robert, and Joseph Tracy. 2021. "All Forecasters Are Not the Same: Time-Varying Predictive Ability across Forecast Environments." Federal Reserve Bank of Cleveland, Working Paper No. 21-06. https://doi.org/10.26509/frbc-wp-202106.

Robert Rich (corresponding author) is at the Federal Reserve Bank of Cleveland (robert.rich@clev.frb.org). Joseph Tracy is at the Federal Reserve Bank of Dallas.
The authors thank James Mitchell, Alexander Chudik, Eric Young, and Edward Knotek for valuable comments. Kristoph Naggert, Victoria Consolvo, and Max Sterman provided excellent research assistance.

I. Introduction

It is widely acknowledged that expectations are important for understanding the decision-making of households and firms, as well as for explaining movements in economic and financial variables. Early work on the formation of beliefs posited that agents form expectations in a static or adaptive manner, but these theories eventually drew criticism because of their restrictions on agents' information sets and allowance for agents to commit systematic forecast errors. In response to these criticisms, the full-information rational expectations (FIRE) model was developed, which assumes that all agents know the true structure of the economy and have access to the same information set.

While the FIRE assumption remains the main paradigm for the formation of expectations, it implies that all agents are identical and therefore cannot generate the type of dispersion in agents' expectations – that is, disagreement – observed in surveys or financial markets. Consequently, in recent years, the FIRE assumption has been replaced with a weaker form of rational expectations in which agents use available information efficiently subject to certain constraints. A prominent feature of these types of models is the presence of informational rigidities (IR) either in the form of sticky information [Mankiw and Reis (2002); Mankiw, Reis, and Wolfers (2003)] or noisy information [Woodford (2002); Sims (2003); Mackowiak and Wiederholt (2009)].

While IR models can generate disagreement at the aggregate level, a key implication of almost all of these models is that heterogeneity in forecast behavior is not persistent at the individual level.[1] That is, while agents can display differences in their forecast behavior at a point in time, their observed behavior over time should on average be the same.[2] In particular, no agent should generate forecasts that are systematically more/less accurate compared to those of other agents.

This paper examines this implication using point and density forecasts from the European Central Bank's Survey of Professional Forecasters (ECB SPF). The survey elicits euro-area expectations for real GDP growth, inflation as measured by the harmonized index of consumer prices (HICP), and the unemployment rate. We analyze the forecast performance of participants to document empirical features that are of general interest for expectations models and have specific relevance for IR models. A central focus of this study is whether participants display comparable

[1] Coibion and Gorodnichenko (2012, 2015) test the predictions of the sticky information model and the noisy information model for various aspects of forecast behavior, including disagreement. However, they conduct the analysis at the aggregate level and do not consider this fundamental implication of the models at the micro level.

[2] The average observed behavior of agents refers to a sufficiently long time period in which individuals either update their information set on a comparable basis in the case of the sticky information model or are subject to a comparable set of shocks in the case of the noisy information model.

forecast accuracy or whether there are systematic differences and, if so, the sources of such differences.

The results provide strong evidence of systematic differences in participants' forecast accuracy. However, the differences are not time invariant, as would be the case if some participants were consistently more/less accurate than other participants. Rather, we find that the rank orderings of participants' accuracy vary with the degree of difficulty of the forecasting environment. That is, some participants display higher relative accuracy in more tranquil environments, while other participants display higher relative accuracy in more volatile environments. In addition, forecast performance is positively correlated across target variables and horizons, with stronger correlation patterns for the density forecast data than for the point forecast data. This latter finding highlights one of the additional insights gained from extending the analysis beyond the point forecast data alone.

Our study makes two contributions to the existing literature on the expectations formation process. The first contribution is the modeling strategy used for the empirical analysis. We adopt a panel data estimation framework in which the individual forecast performance regressions include participant fixed effects and time-specific cross-section averages of forecast performance as explanatory variables, with the latter variable providing a link to the work of Pesaran (2006). Among its many attractive features, the specification allows for a conventional testing procedure for equal forecast accuracy, as well as a flexible approach to control for changes in the degree of difficulty of the forecasting environment. The specification also allows for a richer description of participants' forecast performance and a broader identification of possible sources of heterogeneity.

The second contribution pertains to the features of forecast behavior that we document. Our results are consistent with other work reporting strong evidence of persistent heterogeneity across survey participants in the relative levels of point forecasts, uncertainty, and disagreement.[3] Accordingly, our findings do not support a central prediction of IR models and instead argue for the development of expectations models that not only can generate persistent heterogeneity in key features of forecast behavior but also can account for systematic differences in relative accuracy related to the nature of the forecasting environment.[4]

The paper is organized as follows. The next section provides a summary of the literature evaluating the predictive performance of professional forecasters and testing for the comparability

[3] See Bruine de Bruin et al. (2011), Boero, Smith, and Wallis (2015), and Rich and Tracy (2021).
[4] While this line of research is beyond the scope of the current paper, we view our analysis as providing background and helping to inform the development of such a class of models.

of their accuracy. Section III discusses the modeling strategy and estimation framework used for the empirical analysis. Section IV describes the ECB SPF data. Section V reports the estimation results documenting systematic differences in participants' relative forecast accuracy and how it varies with the difficulty of the forecasting environment. The section also explores other features and properties of participants' forecast accuracy. Section VI concludes by discussing the implications of our findings.

II. Literature Review

This section reviews previous studies testing for differences in forecaster performance. Stekler (1987) examines monthly forecasts from the Blue Chip survey of economic indicators between 1977 and 1982, but restricts the analysis to a balanced panel resulting in 24 forecasters in the sample. While Stekler viewed the goal of identifying the "best" forecasters as problematic, he proposed a methodology that would instead determine whether there were "better" forecasters. His approach calculates the root mean squared error for each year to assign a rank to each forecaster, with the forecaster's rankings then accumulated over the full sample period to produce a rank sum. His analysis indicated that there were significant differences in forecasting ability. Batchelor (1990), however, argued that Stekler's conclusions relied on an incorrectly defined test statistic. When Batchelor applied the appropriate test statistic to the data, he concluded that the evidence did not reject the null hypothesis of equal forecasting ability.

Christensen et al. (2008) examines the US Survey of Professional Forecasters (US SPF) and develops a test for equal forecasting accuracy that draws upon the forecast comparison test of Diebold and Mariano (1995). Their approach, however, is quite restrictive because it requires a balanced panel and a long time series of forecasts. Consequently, they are only able to study three individual forecasters. Their analysis yields mixed results, with tests suggesting equal predictive performance for some variables and not others.

D'Agostino, McQuinn, and Whelan (2012) point out that there are notable drawbacks to the approaches in these previous studies. The requirement of a balanced panel can significantly reduce sample size so that most of the information available from surveys is lost. In addition, metrics based on a period-by-period summation of ranks provide limited information about the nature of participants' forecast errors on either an absolute or a relative basis. To remedy these shortcomings, D'Agostino, McQuinn, and Whelan (2012) develop a test for equal predictive performance that applies bootstrapping and Monte Carlo simulation techniques to a metric based on participants' squared forecast errors. Importantly, their approach recognizes that comparing forecasts within an unbalanced panel can be challenging because of time-variation in the forecasting environment. In

particular, participants generating the same squared prediction error in different periods would not reflect comparable predictive ability if forecasting in some periods is easier/more difficult compared to others. Consequently, the metric used in D'Agostino, McQuinn, and Whelan (2012) involves an adjustment to participants' squared forecast errors.

The approach of D'Agostino, McQuinn, and Whelan (2012) begins by constructing a normalized squared error statistic for each variable for each period for each participant. Abstracting from details related to data and survey features, the normalized squared error statistic for participant *j*, $E^j_{t,t+h}$, is defined as:

$$(1) \qquad E^j_{t,t+h} = \frac{\left(e^j_{t,t+h}\right)^2}{\left(1/N_t\right)\sum_{i=1}^{N_t}\left(e^i_{t,t+h}\right)^2} = \frac{\left(e^j_{t,t+h}\right)^2}{\overline{\left(e_{t,t+h}\right)^2}}$$

where $e^j_{t,t+h}$ is participant *j*'s forecast error associated with the survey prediction in period *t* and the realization of the target variable in period *t+h*, and $\overline{\left(e_{t,t+h}\right)^2}$ is a measure of average forecast performance defined over the N_t survey participants in period *t*. Importantly, the metric in (1) depends on participant *j*'s forecast performance and the performance of the other forecasters. Consequently, the normalization is designed to control for changes in the forecasting environment by generating, for a given value of $\left(e^j_{t,t+h}\right)^2$, a lower (higher) value of $E^j_{t,t+h}$ when forecasters are collectively less (more) accurate compared to periods when they are more (less) accurate.

For each forecaster, an average normalized squared error statistic can be calculated. Letting T^j denote the total number of surveys in which participant *j* appears and T denote the total number of surveys conducted, participant *j*'s average normalized squared error statistic is given by:

$$(2) \qquad \overline{E^j} = \left(1/T^j\right)\sum_{t=1}^{T} E^j_{t,t+h} \ ,$$

where $E^j_{t,t+h}$ is set to zero if participant *j* did not respond to that survey. Because the score in (2) is calculated as an average, it can account for participants entering and exiting the survey.

A historical distribution of forecast performance can be derived using the score in (2) and the associated rank ordering of all participants. The test for equal forecast performance then proceeds by randomly reshuffling and reassigning individual forecasts of a given variable for a

particular survey. The same procedure is applied to each survey, resulting in a new sequence of forecasts for each participant that can be used to calculate an overall score from (2) and to construct a rank ordering. The process is repeated many times to generate a large number of simulated distributions of forecaster performance. The test for equal forecast performance compares the historical distribution of forecast performance to the simulated distributions. Under the null hypothesis of equal forecaster ability, the historical distribution of forecast performance should lie within selected percentiles of the simulated distribution that serve as confidence intervals. Segments of the historical distribution that lie below (above) the lower (upper) bootstrap confidence intervals are indicative of forecasters who are significantly better (worse) than their peers.

D'Agostino, McQuinn, and Whelan (2012) apply their testing procedure to data from the US SPF. The sample period begins in 1968:Q4 and ends in 2009:Q3, with 309 forecasters appearing over the time period. Because participation rates vary among the forecasters, the analysis uses the full sample of forecasters and a restricted sample that excludes forecasters who provide fewer than 10 forecasts. The bootstrap distributions involve 1,000 simulations and the 1st and 99th bootstrap percentiles serve as confidence bands. The results suggest that most forecasters display equal predictive performance, although there is evidence that a relatively small group of forecasters perform very poorly.

Meyler (2020) examines the issue of equal forecasting ability for participants in the ECB SPF. As background, he notes that the testing procedure of D'Agostino, McQuinn, and Whelan (2012) relies on participants' forecast errors being uncorrelated across periods. When the data in (1) involve overlapping forecast horizons ($h > 1$), the conventional application of the testing procedure is not valid because of autocorrelation in the forecast errors. To remedy this situation, Meyler (2020) proposes separating the data across nonoverlapping forecast horizons.

Meyler (2020) applies the bootstrapping and Monte Carlo simulation techniques of D'Agostino, McQuinn, and Whelan (2012) to quarterly one-year-ahead and one-year/one-year-forward forecasts from the ECB SPF during the period 1999-2018. Given the structure of the survey instrument, the correction for autocorrelation involves grouping SPF rounds that are four quarters apart. To avoid the influence of forecasters with very limited participation, Meyler (2020) excludes forecasters who provide fewer than 20 forecasts for a particular target variable and horizon, resulting in the number of forecasters in the sample varying between 63 and 77 over the time period.[5] He also performs a robustness check using pre- and post-global financial crisis samples. For

[5] As discussed shortly, Meyler (2020) applies his participation restriction to the full time period and not to the four quarterly sub-samples.

the analysis, the bootstrap distributions involve 1,000 simulations and the 1st and 99th bootstrap percentiles serve as confidence bands. The results provide little evidence of forecasters who perform significantly better or worse than their peers.

The approach adopted by D'Agostino, McQuinn, and Whelan (2012) and Meyler (2020) is attractive for several reasons. As previously discussed, the metric in (1) controls for changes in the degree of difficulty in forecasting a variable over time and the score in (2) can account for the varying entry and exit of survey participants. Meyler (2020) also points out that the approach allows for the aggregation across target variables and horizons. There are, however, several issues pertaining to the methodology and data in D'Agostino, McQuinn, and Whelan (2012) and Meyler (2020) that appear problematic.

One issue is related to the overlapping forecast horizons of the data. While the modification proposed in Meyler (2020) controls for autocorrelation in the forecast errors, it results in a significant loss of efficiency because of the reduced time series dimension of the data. Even if a respondent were to participate in every survey from 1999:Q1 through 2018:Q4, restricting the analysis to SPF rounds four quarters apart only allows for a maximum of 20 observations to compute the relevant score in (2) for each sub-sample associated with a target variable and horizon. The number of observations, however, will typically be lower for most respondents because they will not be full-time participants.

Another issue concerns the metric in (1) to control for changes in the forecasting environment. While the normalization scheme may seem reasonable, other approaches are available to capture the effect of a common element leading most participants to display higher or lower accuracy in a period. For example, time fixed effects is a common approach used to control for aggregate shocks in a panel data setting.[6] It is important, therefore, to investigate the appropriateness of the normalization scheme and to consider alternative modeling strategies that can serve the same purpose. In addition, Meyler (2020) points out that a drawback of the normalized metric in (1) is that it involves an asymmetric treatment of accuracy at the individual level versus the aggregate level. Specifically, a forecaster who makes a relatively large error when forecast errors on average are small will incur a large penalty, whereas a forecaster who makes a relatively small error when forecast errors on average are large will not benefit much.

A final issue is that D'Agostino, McQuinn, and Whelan (2012) and Meyler (2020) restrict their attention to point forecasts. In the case of the US SPF and the ECB SPF, however, matched density forecasts are also available to test for equal predictive ability across forecasters. Taken

[6] We discuss the issue of modeling a common element across forecasters in more detail in the next section.

together, these issues raise questions about the reliability and robustness of the conclusions in D'Agostino, McQuinn, and Whelan (2012) and Meyler (2020) and motivate our study.

III. An Empirical Framework to Test for Equal Forecast Accuracy

In contrast to other studies, we propose a heterogeneous panel data model to test the hypothesis of equal forecast accuracy across survey participants. Specifically, the empirical analysis is based on the following specification for each participant *j* and survey in period *t*:

$$FP^{j}_{t,t+h} = \alpha_j + \lambda_j \left(\overline{FP_{t,t+h}} \right) + \varepsilon^{j}_{t,t+h} \tag{3}$$

where $FP^{j}_{t,t+h}$ and $\overline{FP_{t,t+h}}$ denote a forecast performance (FP) metric at the individual and average (cross-section) level, respectively, and $\varepsilon^{j}_{t,t+h}$ is the error term. For the moment, we only note that higher (lower) values of the FP and $\overline{FP}$ measures denote lower (higher) forecast accuracy and defer a more detailed discussion of the specific metrics until the next section.

There is a close parallel between the variables and structure in (3) and those in (1). On a general level, the panel data model can be viewed as a regression-based analogue to the methodology used in D'Agostino, McQuinn, and Whelan (2012) and Meyler (2020). As we argue, however, there are several advantages to our empirical framework. First, the participant-specific intercepts and slopes allow for a deeper exploration into the issue of heterogeneity. In particular, we can evaluate individual forecast performance through two channels: a participant fixed effect α and a time-varying component $\lambda\left(\overline{FP}\right)$ related to the average forecast performance of participants in a survey. The linear specification in (3) also addresses Meyler's (2020) criticism about the normalized metric in (1) and its asymmetric treatment of forecast accuracy at the individual level versus the aggregate level.

The specification in (3) also lends itself to a convenient and straightforward test of equal forecast accuracy through the joint restriction $\alpha_j = 0$ and $\lambda_j = 1$ across all survey participants. The testing procedure formalizes the idea that if participants are identical on average, then their observed behavior should be similar and individuals should not display forecasts that are systematically more/less accurate compared to those of other participants. That is, everyone should look indistinguishable from each other over time. Importantly, the specification in (3) also nests two alternative approaches to control for changes in the forecastability of a variable. For example, the normalization procedure adopted by D'Agostino, McQuinn, Whelan (2012) and Meyler (2020)

corresponds to the special case of the α_j's being jointly equal to zero. Alternatively, the application of time fixed effects corresponds to the case of equality of the λ_j's. Consequently, we have the additional capability to evaluate specific approaches to address the issue of variation in the forecasting environment.

Another point of consideration is the important role of $\overline{FP}$ in our analysis. A typical concern that emerges in estimating panel data models is the cross-sectional correlation across units. Estimators that fail to account for cross-sectional dependence turn out to be inefficient and inconsistent [Sarafidis and Wansbeek (2012)]. Our modeling strategy allows us to interpret $\overline{FP}$ as cross-section averages of dependent variables and to view this regressor within the context of the common-correlated effects (CCE) procedure of Pesaran (2006). This procedure not only accounts for dependence across participants but also provides the basis to view $\overline{FP}$ as a proxy for an unobserved common factor, such as an aggregate shock, that impacts the forecasting environment and generates higher or lower accuracy across participants in a period. Accordingly, the movements in $\overline{FP}$ are a very natural way to describe time variation in the forecasting environment.

Another advantage of our estimation framework is that it can account for conditional heteroscedasticity and autocorrelation in the model disturbance terms. As previously discussed, the latter issue arises when the data involve overlapping forecast horizons (h >1) and is problematic for the approach adopted by D'Agostino, McQuinn, and Whelan (2012) and Meyler (2020). In contrast, we can apply the Newey-West (1987) covariance matrix modified for use in a panel data setting to obtain robust standard errors for the estimated parameters in (3). Consequently, we do not need to separate the panel into survey rounds at a specified distance apart as in Meyler (2020), but instead we can exploit the efficiency gains from using all of the information on participants in a collective manner.

Our modeling strategy also lends itself to making interpersonal and intrapersonal comparisons. Because of the nature of the regression equations in (3), it is useful to note that the estimated values for α and λ across participants will be centered around 0 and 1, respectively. By partitioning the parameter space into four quadrants around these means, as shown in Figure 1, we can examine the relationship between the estimated parameter pairings $\left(\hat{\alpha}_j, \hat{\lambda}_j\right)$ for evidence of patterns in participants' forecast behavior.

If we were to reject the null hypothesis of equal predictive performance, then one possibility for this outcome is that a scatterplot of the estimated parameter pairings in Figure 1 would

principally run from the $(\alpha < 0, \lambda < 1)$ quadrant up through the $(\alpha > 0, \lambda > 1)$ quadrant. In this hypothetical configuration, the former quadrant would indicate there are participants who are more accurate on average than their peers irrespective of the forecasting environment, while the latter quadrant would indicate there are participants who are less accurate on average than their peers irrespective of the forecasting environment.[7] There is, however, a second possible configuration where the estimated parameter pairings principally run from the $(\alpha < 0, \lambda > 1)$ quadrant down through the $(\alpha > 0, \lambda < 1)$ quadrant, implying that relative forecast performance varies with changes in the forecasting environment. Specifically, participants in the $(\alpha < 0, \lambda > 1)$ quadrant are relatively more accurate in a tranquil environment and then relatively less accurate as the environment becomes more volatile. The opposite case holds for participants in the $(\alpha > 0, \lambda < 1)$ quadrant. Inspection of the scatterplot can also be informative about the dispersion of the estimated parameter pairings.

Estimation of (3) also allows us to generate a forecast performance profile for each participant based on the predicted values of the model $\left(\widehat{FP}\right)$ where:

$$\widehat{FP}^{j}_{t,t+h} = \hat{\alpha}_j + \hat{\lambda}_j \left(\overline{FP_{t,t+h}}\right) \tag{4}$$

The behavior of the performance profiles depends on the nature of the forecasting environment as well as the quadrants of the parameter space. As we vary the value of $\overline{FP}$, the performance profile of participants located in the $(\alpha < 0, \lambda < 1)$ quadrant will run below the profile of those located in the $(\alpha > 0, \lambda > 1)$ quadrant. In the case of participants in the $(\alpha < 0, \lambda > 1)$ quadrant and $(\alpha > 0, \lambda < 1)$ quadrant, however, their profiles will display crossings that reflect changes in the rank orderings of forecast accuracy that vary with the forecasting environment.

Our earlier discussion of IR models highlighted the absence of a mechanism to generate persistent heterogeneity in forecast accuracy for a given target variable and horizon. However, we can extend the analysis to consider other aspects of forecast behavior. Specifically, IR models would also imply that the estimated parameter pairing or the predictive performance of a participant should not display systematic patterns over time. We can also use our empirical framework to investigate

[7] Recall from the previous discussion that higher (lower) values of the individual FP measures are associated with less (more) accurate forecasts.

these implications. For example, information on $\left(\hat{\alpha}_j, \hat{\lambda}_j\right)$ can determine the quadrant location of a participant. In addition, we can calculate the following average relative forecast performance metric:

$$(5) \qquad \left(\overline{FP^j - \overline{FP}}\right) = \left(1/T^j\right) \sum_{t=1}^{T} \left(FP_{t,t+h}^j - \overline{FP_{t,t+h}}\right)$$

where $FP_{t,t+h}^j$ is set to $\overline{FP_{t,t+h}}$ if participant *j* did not respond to that survey. The metric in (5) is similar to the score previously described in (2) and provides an assessment of a participant's overall forecast performance. Evidence that a participant's estimated parameter pairings tend to locate in the same quadrant or evidence of a positive association between metrics from (5) across target variables and horizons would indicate that a participant's forecast behavior displays similar features.[8]

Taken together, our modeling strategy provides a unified empirical framework to analyze the forecast behavior of ECB SPF participants and inform models of the expectations formation process. With regard to IR models, our approach affords several avenues by which to gauge whether there are systematic differences across participants as well as similar behaviors for an individual participant. Moreover, our approach uses conventional estimation and testing procedures and also accounts for a range of econometric issues arising from the nature of the survey instrument and data. Importantly, changes in the degree of difficulty in forecasting a variable does not present a challenge to our study or require the adoption of some type of split sample analysis. Rather, time variation in the forecasting environment is an integral element in our methodology and plays a central role in our ability to compare and contrast various features of participants' predictive ability.

IV. The European Central Bank Survey of Professional Forecasters

The ECB SPF began in January 1999 and provides a quarterly survey of forecasts for the euro area. The survey draws its pool of panelists from both financial and nonfinancial institutions, with most, but not all, located in the euro area. Meyler (2020) notes that the principal aim of the survey is to solicit expectations about inflation, real GDP growth, and unemployment, although the questionnaire also contains a noncompulsory section asking participants for their expectations of other variables and to provide qualitative comments that inform their quantitative forecasts.[9] The ECB SPF asks panelists for forecasts at short-, medium- and longer-term horizons, including both "rolling" and "calendar year" variants. The survey is fielded in February, May, August, and

[8] While the empirical analysis in Section V restricts its attention to pairwise comparisons of participants' forecast accuracy across either different target variables/same horizon or different horizons/same target variables, we can extend the dimensions under consideration.

[9] The additional expectations pertain to variables such as wage growth, the price of oil, and the exchange rate.

November, with a little under 50 panelists on average responding per survey. For additional details about the ECB SPF, see Garcia (2003) and Bowles et al. (2007).

We examine forecasts for real GDP growth, inflation as measured by the harmonized index of consumer prices (HICP), and the unemployment rate. This choice partly reflects the structure of the survey instrument that asks respondents to submit both point- and density-based forecasts for these three macroeconomic variables.[10] Because D'Agostino, McQuinn, and Whelan (2012) and Meyler (2020) restrict their analyses to point-based forecasts, the inclusion of both types of forecasts offers an important robustness check. For the density forecasts, participants report their subjective probability distribution of forecasted outcomes as a histogram using a set of intervals provided in the survey. While the ECB SPF occasionally changes the number of closed intervals for the histogram, it has essentially maintained a common bin width for the closed intervals throughout its history.[11]

With regard to forecast horizons, we examine point and density forecasts that involve rolling one-year-ahead and one-year/one-year-forward horizons. Compared to calendar year horizons, an advantage of the rolling horizons is that the horizon length remains constant through time and allows us to treat the data as quarterly observations on a set of homogeneous series. As Garcia (2003) notes, there is a temporal misalignment between the target variables because of differences in the data frequency and publication lags of the variables. Specifically, real GDP growth is published quarterly with a two-quarter lag, while HICP inflation and the unemployment rate are published monthly with a one-month lag and a two-month lag, respectively.[12]

Our study analyzes surveys conducted from 1999:Q1–2018:Q3, with forecast evaluation ending in 2019:Q3 for all series. The ECB SPF, like other surveys, has experienced entry and exit of respondents over time. Moreover, occasionally participants do not respond to the complete questionnaire or to individual items within the questionnaire. As noted by Meyler (2020), participants provide the highest number of forecasts for HICP inflation and the lowest number for unemployment, with the number of forecasts at the one-year-ahead horizon exceeding that at the

[10] The ECB SPF is among a small, but growing number of surveys that solicit both point and density forecasts. Other notable surveys include the US SPF (published by the Federal Reserve Bank of Philadelphia), the Bank of England Survey of External Forecasters, and Federal Reserve Bank of New York Survey of Consumer Expectations.

[11] The only deviation in this design started with the 2020Q2 survey in response to the COVID-19 outbreak. The (nearly) constant interval width of the ECB SPF density forecasts is in contrast to the US SPF density forecasts, which have experienced periodic changes in interval widths.

[12] For example, the 2010Q1 survey questionnaire asks respondents to forecast one-year-ahead output growth from 2009Q3–2010Q3. For HICP inflation, the corresponding forecast horizon is December 2009–December 2010. For the unemployment rate, the corresponding forecast is for November 2010.

one-year/one-year-forward horizon. Participants also report more point forecasts than density forecasts. Given the unbalanced panel structure of the ECB SPF, we only include participants at each individual target variable/horizon who provide at least 50 forecasts. Further, we only consider matched point and density forecasts to maintain comparability across the types of forecasts.[13] Consequently, the number of participants varies from 34 (HICP inflation at the one-year-ahead horizon) to 21 (unemployment rate at the one-year/one-year-forward horizon). Compared to the number in Meyler's (2020) study, the number of participants in our panel is smaller, although the number of recorded forecasts for each participant is notably higher.[14]

An extremely important consideration for the analysis is the issue of data vintage to construct realizations of the target variables. As is the case for most macroeconomic data for most countries, euro-area macroeconomic statistics tend to be revised from preliminary releases. Consequently, a choice must be made about the relevant release associated with a participant's forecast. Following Meyler (2020), we construct realizations of the target variables for HICP inflation and the unemployment rate using monthly data from the first full release.[15] For real GDP growth, we construct realizations of the target variables using quarterly data from the second estimate. We have considered other approaches to construct realizations of the target variables as additional robustness checks.[16]

The ability to assess a participant's forecast performance requires not only the choice of data vintages but also measures of point and density forecast accuracy. For the point forecasts, we adopt the absolute error as the metric:

$$(6) \qquad {}^{POINT}FP^{j}_{t,t+h} = \left| X_{t+h} - E^{j}_{t}\left[X_{t+h} \right] \right|$$

where X_{t+h} denotes the realized value of the relevant ECB SPF target variable in period $t+h$ and $E^{j}_{t}\left[X_{t+h} \right]$ denotes the reported point forecast from participant j in the survey at date t.

For the density-based accuracy measure, we adopt the absolute rank probability score (ARPS) as the metric:

---

[13] We also exclude any participant whose probabilities for a density forecast do not sum to unity.

[14] Recall from the earlier discussion that Meyler divides the forecasts into separate quarterly rounds. We have also experimented with a lower threshold of 40 participants and obtained similar results.

[15] Specifically, if the target variable is one-year-ahead HICP inflation, we use the first full release reporting the value of the price index in month $t+12$. The same release is used to obtain the value of the price index in month t.

[16] We used current vintage data as one robustness check. As another robustness check, we construct growth rates using the first full release to obtain the value of the price index in month t and the second estimate for the level of real GDP in quarter t. The results changed very little using these alternative approaches.

$$(7) \qquad {}^{DENSITY}FP^{j}_{t,t+h} = \frac{1}{k_t - 1}\sum_{i=1}^{k_t}\left|\sum_{g=1}^{i} {}_g p_t^j - \sum_{g=1}^{i} {}_g I_{t+h}\right|$$

where we assume there are k_t bins associated with the histogram for the survey at date *t*, the probability assigned by respondent *j* to the g^{th} bin is ${}_g p_t^j$, and ${}_g I_{t+h}$ denotes an indicator variable that takes a value of one if the actual outcome in period *t+h* is in the g^{th} interval of the histogram from the survey at date *t*. The ARPS has the property that a participant receives "credit" by assigning probability in bins close to the bin containing the actual outcome.[17]

The evaluation of the ECB SPF density forecasts requires additional discussion beyond the selection of a metric. To the extent that respondents place any probability in either open interval, the manner chosen to close off the open intervals will affect the value of the forecast performance metric in (7). We follow a common—although ad hoc—assumption and close the exterior open intervals by assigning them twice the width of the interior closed intervals. We also need to address the issue of the location of probability mass associated with the density forecasts. We again draw upon common practices and assume that the probability mass is distributed uniformly within each bin of the histogram. Finally, we exclude the 2009:Q1 one-year-ahead real GDP growth density forecast data because many respondents placed significant probability in the lower open interval of the histogram in this survey.[18]

V. Empirical Results

It is instructive to begin by examining the behavior of the average forecast performance metrics $\left(\overline{FP}\right)$ to compare forecastability across the variables as well as to identify low- and high variance episodes. Figure 2 plots the movements of $\left(\overline{FP}\right)$ for the point forecasts and density

[17] The squared error metric used in D'Agostino, McQuinn, and Whelan (2012) and Meyler (2020) provides an alternative to the absolute value norm in (6) and (7). However, measures based on the squared norm are more sensitive to outliers compared to measures based on the absolute value norm. Our subsequent discussion also points out that predictive performance can be sensitive to the treatment of open intervals of the density forecasts, with the sensitivity again being higher when using a squared norm. Because the absolute value norm helps to mitigate both of these concerns, we adopt this metric for the distance calculations in the paper. However, we used the squared norm for the point forecasts and density forecasts for robustness and found similar results. These results are available upon request.

[18] For this survey, the significant probability mass at the lower open interval corresponded to a growth rate of "-1 percent or less" and was due to the survey design of the density forecasts and its inability to provide sufficient coverage for the pessimistic point predictions of output growth. For individuals who either reported point predictions below -1 percent or wanted to indicate significant downside risk, they assigned most of their probability to the open-ended interval. See Abel et al. (2016) for further discussion.

forecasts of real GDP growth, HICP inflation, and the unemployment rate at the one-year-ahead horizon, while Figure 3 provides the corresponding information at the one-year/one-year-forward horizon. The metrics are plotted based on the realization of the target variable, with gray bars indicating recessions as determined by the Euro Area Business Cycle Dating Committee of the Center for Economic Policy Research.[19]

As shown, there is generally a close correspondence between the point and density forecast performance metrics for the same target variable and horizon. While the difficulty of forecasting outcomes around the time of the global financial crisis and the euro-area debt crisis is evident, the data speak to other episodes associated with sizable forecast errors that are not uniform in their timing across the three target variables. Consequently, there is sufficient variability in the forecasting environments to mitigate concerns that our results are largely driven by just a few events. In terms of the pattern of the forecast errors, they are highest for real GDP growth around the time of the global financial crisis. For HICP inflation, they are highest around the time of the global financial crisis as well but are also elevated toward the beginning of the sample and during the middle of the last decade. For the unemployment rate, the forecast errors are again largest around the time of the global financial crisis, although they are also elevated at the beginning of the sample and around the time of the euro-area debt crisis.

Table 1 and Table 2 present formal tests for equal predictive ability using the point forecasts and density forecasts, respectively. We estimate the parameters in (3) using ordinary least squares (OLS), with standard errors computed using the Newey-West (1987) covariance matrix estimator modified for use in a panel data set.[20] Letting $\hat{\theta} = \left[\left(\hat{\alpha}_1, \hat{\lambda}_1\right), \left(\hat{\alpha}_2, \hat{\lambda}_2\right), ..., \left(\hat{\alpha}_N, \hat{\lambda}_N\right)\right]$ denote the vector of estimated parameters of the model, we construct the following Wald test statistic for the joint null hypothesis that $\alpha_i = 0$ and $\lambda_i = 1$ for all i =1,…,N participants included in the panel for a particular target variable and horizon:

$$W = \left(\hat{\theta} - \theta_0\right)' \left[\text{var}^{-1}\left(\hat{\theta}\right)\right] \left(\hat{\theta} - \theta_0\right) \tag{8}$$

[19] For example, the metric associated with the forecasts of HICP inflation from 2015Q1-2016Q1 is plotted at 2016Q1. As a reminder, we plot the $\left(\overline{FP}\right)$ value for the 2009:Q1 one-year-ahead point forecasts of real GDP growth in Figure 2 but do not include these data in the analysis owing to the exclusion of the matched density forecasts. Unlike the absolute error metric, the ARPS metric is restricted to fall in the range between 0 and 1.

[20] We allow the error terms to follow a fourth-order moving average process to account for the overlap of forecast horizons.

Because our estimation approach nests two common approaches to control for variability in the forecasting environment, we can also use the estimated parameter vector $\hat{\theta}$ and the estimated variance-covariance matrix $\text{var}(\hat{\theta})$ to construct Wald test statistics for the validity of the normalization approach $(\alpha_1 = \alpha_2 = \ldots = \alpha_N = 0)$ or the use of time fixed effects $(\lambda_1 = \ldots = \lambda_N)$.

As shown by the values of the test statistic, we strongly reject the hypothesis of equal predictive ability in all cases except for the point forecasts of inflation at the one-year horizon.[21] In terms of controlling for changes in the forecast environment, we reject the normalization approach and the use of time fixed effects in all cases except for the point forecasts of inflation at the one-year horizon. Consequently, while the normalization approach used in D'Agostino, McQuinn, and Whelan. (2012) and Meyler (2020) may have intuitive appeal and plays a key role in their analyses, the data do not support this approach.

To gain insight into the sources for the rejection of equal predictive ability, Figure 4 and Figure 5 display scatterplots of the individual estimated parameter pairings $(\hat{\alpha}_j, \hat{\lambda}_j)$ for the point forecasts and density forecasts, respectively. The patterns are striking in their similarity across target variables and horizons as well as for the point and density forecasts. The estimated parameter pairings do not support the idea that, looking across all forecasting environments, there are some participants who are typically more accurate and other participants who are typically less accurate. Rather, the estimated parameter pairings display a strong inverse relationship consistent with the view that participants' relative predictive performance varies with the forecasting environment. Moreover, the patterns do not indicate clustering and do not suggest that the inverse association reflects the behavior of a few participants. Rather, the observations largely fall in the $(\alpha < 0, \lambda > 1)$ and $(\alpha > 0, \lambda < 1)$ quadrants at a similar frequency and are fairly disperse within each quadrant.

There are, however, two notable differences for the estimated parameter pairings across the point and density forecasts. First, the slopes of the linear relationships between the estimated parameter pairings are typically steeper for the density forecasts as compared to the point forecasts. As we discuss in more detail shortly, this implies that the "crossing point" for the forecast performance profiles of participants in the $(\alpha < 0, \lambda > 1)$ and $(\alpha > 0, \lambda < 1)$ quadrants – that is, the level of average forecast performance where there is a switch in their relative accuracy – is lower for density forecasts.

[21] The different degrees of freedom reflect the varying number of respondents meeting the participation restriction for the various target variables and horizons.

A second difference concerns the relative precision of the estimated parameter pairings. To illustrate this feature of the data, we adopt Monte Carlo simulation techniques and generate 1,000 draws of the parameter pairings vector for each target variable and horizon using the estimated joint normal distribution for $\hat{\theta} = \left[\left(\hat{\alpha}_1, \hat{\lambda}_1\right), \left(\hat{\alpha}_2, \hat{\lambda}_2\right), ..., \left(\hat{\alpha}_N, \hat{\lambda}_N\right)\right]$. Figure 6 and Figure 7 plot the distributions for the point forecasts and density forecasts, respectively, where the estimated pairings are in black and the simulated pairings are in gray. As shown, the distributions for the density forecasts are much tighter as compared to the point forecasts. The differences in the degree of precision are notable and consistent with the evidence in Tables 1 and 2 that the rejections of the various hypotheses are generally stronger for the density forecast data. We will also explore how these differences bear upon issues related to the quadrant location of a participant.

The evidence in Tables 1 and 2 and in Figures 4 and 5 suggests two principal reasons why our conclusions differ from those of D'Agostino, McQuinn, and Whelan (2012) and Meyler (2020). In their approach, D'Agostino, McQuinn, and Whelan (2012) and Meyler (2020) only relate predictive performance to average forecast performance and impose the same inverse weighting scheme across participants. However, we find notable and statistically significant variability across participants' fixed effects and slope coefficients that speaks to the importance of both dimensions as well as parameter flexibility to characterize predictive performance. The broader and more general nature of our approach affords greater scope to identify differential behavior among forecasters and gauge the persistence of the heterogeneity.

Drawing upon our earlier discussion, we can examine the implications of the estimation results for the forecast performance profiles of participants. To do so, we select a participant from each of the four quadrants associated with a target variable and horizon. While any scatterplot can be used for the exercise, it is convenient to use the one-year-ahead point forecasts of the unemployment rate for this illustration. Using the estimated parameter pairing for each participant identified by the red circles in the relevant panel in Figure 4, we vary the average forecast performance metric $\left(\overline{FP}\right)$ and trace out the corresponding forecast performance profiles.

Figure 8 depicts the forecast performance profiles of the four participants that closely align with our expected behaviors. The performance profiles of the participants in the $(\alpha < 0, \lambda < 1)$ and $(\alpha > 0, \lambda > 1)$ quadrants, respectively, are typically the lowest and highest, and then steadily widen as the forecast environment transitions from low to high variance. In contrast, the performance profiles of the participants in the $(\alpha > 0, \lambda < 1)$ and $(\alpha < 0, \lambda > 1)$ quadrants cross at 0.435, which

corresponds to the 40th percentile of the $\overline{FP}$ values. At values below the crossing, the participant in the $(\alpha < 0, \lambda > 1)$ quadrant displays greater relative forecast accuracy, while the opposite holds for the participant in the $(\alpha > 0, \lambda < 1)$ quadrant at values above the crossing.[22]

The performance profiles offer a visual perspective into our methodology and findings. As shown, a key aspect of our study is the ability to discern and categorize divergent forecast accuracy, which is especially important owing to the higher incidence of participants in the $(\alpha < 0, \lambda > 1)$ and $(\alpha > 0, \lambda < 1)$ quadrants. Consequently, we expect the rank orderings of participants to display considerable variability as the degree of difficulty of the forecasting environment changes.

To gain a better appreciation of the extent of the variability, we now consider all of the participants associated with the one-year-ahead point forecasts of unemployment. Specifically, we construct rank orderings based on participants' forecast accuracy evaluated at five equally spaced values spanning the range of $\overline{FP}$ for this target variable and horizon. Table 3 reports the results. To facilitate the discussion, we list the participants based on their rankings at the lowest selected value of $\overline{FP}$. We can then examine how the rankings change as the forecasting environment becomes more difficult. As shown, the pattern of the rank orderings is consistent with the evidence from the scatterplot of the estimated parameter pairings in the lower left panel of Figure 4. Some individuals largely maintain the same rankings either because they tend to be highly accurate (#96), highly inaccurate (#36), or close to the average most of the time (#16). There are also examples of dramatic improvements (#91, #52) and dramatic declines (#1, #2) in rankings as the forecasting environment changes from easy to difficult. However, it is more typical to observe individuals who, as the forecasting environment turns more challenging, become relatively more accurate (#15, #98) or relatively less accurate (#4, #89), resulting in rank orderings that are predominantly marked by crossings.

---

[22] It is important to recall that the quadrants characterize a participant's forecast accuracy relative to the (cross-section) average. For example, participants in the $(\alpha < 0, \lambda < 1)$ quadrant display forecasts that are systematically more accurate than the average. However, this does not imply that these participants' forecasts always outperform those of individuals in other quadrants. Consequently, there is no inconsistency with the figure displaying an additional crossing at 0.781 in which the participant in the $(\alpha > 0, \lambda < 1)$ quadrant begins to display greater relative forecast accuracy compared to the participant in the $(\alpha < 0, \lambda < 1)$ quadrant. The crossing point corresponds to the 84th percentile of the $\overline{FP}$ values and is associated with a high-variance episode, which is the type of forecasting environment in which participants in the $(\alpha > 0, \lambda < 1)$ quadrant display enhanced performance.

The analysis up to this point has examined forecast data for the target variables and horizons in isolation. However, there is scope to investigate if there are similarities in the data across target variables and horizons. For example, while the scatterplots show that the estimated parameter pairings principally lie in the $(\alpha < 0, \lambda > 1)$ and $(\alpha > 0, \lambda < 1)$ quadrants, they do not indicate the extent to which the pairings for a participant tend to locate in the same quadrant. Another consideration is the extent to which a participant's predictive performance for a target variable and horizon correlates with performance for other target variables and horizons. These additional features are informative for the development or evaluation of expectations models where, in the latter case, our previous discussion has noted the relevance for IR models.

Our investigation into the location of the parameter pairings for a participant requires more than simply focusing on the quadrant associated with the estimates. We must also account for the uncertainty associated with the estimates. Consequently, we use the distributions in Figures 6 and 7 generated from 1,000 simulations and calculate the aggregate percentage of parameter pairings located in each quadrant based on all relevant simulations for a participant. Figure 9 and Figure 10 report the highest aggregate percentage in a quadrant for a participant's point forecasts and density forecasts, respectively. The values are the ratio of the combined number of simulated parameter pairings for a participant that fall in each quadrant to the total of 6,000 simulations. For purposes of comparison, we present the results for the 23 participants included in all six combinations of target variables and horizons.[23] Using 50 percent as an arbitrary threshold, the evidence in Figure 9 shows only modest support for the idea that a participant's parameter pairings tend to locate in the same quadrant. Specifically, the histogram bars show that only a third of the participants exceed the threshold criterion. For those participants who exceed the threshold criterion, the percentages generally are not significantly higher than the 50 percent value.

A very different picture emerges when we look at the density forecasts. There are now 19 participants who exceed the threshold criterion and the calculated percentages are notably higher than the 50 percent value in many cases. Particularly noteworthy are the two participants whose forecast behavior suggests almost no affiliation with the other three quadrants across the six combinations of target variables and horizons. The results in Figure 9 and Figure 10 are consistent with the evidence from Figure 7 and are another example of the different conclusions that can be drawn between the point and density forecasts. Moreover, the two types of data lead to a very different narrative describing the connection between a participant's predictive performance and the

[23] We have also extended the analysis to include the other participants in our study and the results are similar.

nature of the forecasting environment, with the density forecasts indicating that there is considerably more overlap.

We can also use the average relative forecast performance metrics in (5) to make various comparisons across the forecast data, where lower values again indicate better predictive performance. Because of the large number of comparisons, we only provide a summary of the results. Overall, we find that forecast performance correlates positively across horizons and target variables in almost all cases. There are, however, differences across various dimensions that are worth noting. One difference is a continuation of earlier discussions and concerns the type of forecast data. Specifically, we find that the density forecast data generate a much stronger association than the comparable point forecast data. The top panel in Figure 11 is representative of this finding and shows scatterplots of the average relative forecast performance metrics for inflation at the two forecast horizons for the point forecast data and density forecast data, respectively. While the point forecast data indicate a modest correlation of 0.43, the correlation of 0.77 for the density forecast is nearly twice as high. Looking across all of the pairwise combinations of target variables and horizons, the correlations for the point forecast data are typically in the 0.2-0.4 range, while the correlations for the density forecast data are in the 0.7-0.8 range.

Another feature of forecast performance that emerges is that the correlations are generally higher for the same target variable at different horizons than for different target variables at the same horizon. An ordered ranking of the correlations indicates that the lowest three values are associated with inflation and GDP growth at the two forecast horizons and GDP growth and unemployment at the one-year/one-year-forward horizon. In contrast, the highest three values are associated with unemployment (using both types of forecast data) and inflation at the two forecast horizons. While a more detailed analysis is beyond the scope of the current paper, such behavior could reflect some type of heterogeneity in participants' loss functions or processing capacity that is in turn related to information acquisition or signal monitoring.

A further examination of forecast performance across the target variables reveals two other features. First, there tends to be a stronger correlation at the shorter horizon. The middle panel of Figure 11 shows scatterplots of the average relative forecast performance metrics for GDP growth and unemployment. Unlike the pattern at the one-year-ahead horizon, there is much less of a translation of forecast performance from unemployment into GDP growth at the one-year/one-year-forward horizon. Second, there is less of a linkage between forecast performance for GDP growth and inflation than there is for GDP growth and unemployment. The bottom panel of Figure 11 shows the scatterplots of the corresponding average relative forecast performance metrics at the

one-year-ahead horizon. For inflation and GDP growth, forecast performance actually shows a slight inverse relationship.[24] In the case of GDP growth and unemployment, however, there is a meaningful positive association.

VI. Conclusion

This paper develops an alternative empirical framework to investigate whether ECB SPF participants display systematic differences in their predictive performance. While the nature of the accuracy of professional forecasters is of interest by itself, our study draws further motivation from IR models and the implication that the forecast data should not display evidence of persistent heterogeneity. In addition to making comparisons of accuracy across participants, we investigate the correlation patterns of participants' accuracy across parameter configurations, target variables, and horizons. As a robustness check, we also consider the evidence from density forecasts, which represent a data source that has been largely overlooked in previous analyses.

Based on forecasts for output, inflation, and unemployment, we find strong evidence of systematic differences in participants' accuracy that are not related to innate ability but instead are episodic in nature. By way of a simple illustration, the results suggest that participants largely divide into two "camps": those who display relatively better forecasting performance in low-variance times and those who do so in high-variance times. Consistent with this view, we find considerable variability in the rank orderings of participants. The analysis from examining the data across parameter quadrants, target variables, and horizons sheds light on additional features that help to enhance the forecast profile of participants, although the results can be sensitive to the type of forecast data. Compared to the point forecast data, the density forecast data indicate marked similarities in the forecast behavior of participants. In terms of the illustration introduced above, participants tend to locate in the same "camp" across target variables and horizons. There is also a positive and meaningful relationship between the forecast performances of a target variable across different horizons. On balance, we conclude that participants do not display comparable forecast behavior and that observed differences are not due to random variation.

It is evident that heterogeneity is a prominent feature of expectations data. While various models of expectations formation have been successful at generating differential forecast behavior, they typically are unable to account for the systematic nature of the differences documented in this study. It would be interesting and important to determine if these same empirical features are

[24] This is the only instance where the average relative forecast performance metrics display a negative relationship.

present in other survey data.[25] If so, then the findings would argue further for the development of expectations models that can generate persistent heterogeneity in key features of forecasters' behavior, but can also account for the differential effects of the forecast environment on predictive performance. The opportunity to explore and identify the key underpinnings in the course of such a development would serve as fertile ground for future research.

[25] The US SPF would seem to be a natural candidate for such an investigation. It is unclear, however, if a parallel analysis can be conducted for the US SPF because of differences in the survey instrument. Specifically, the US SPF does not feature "rolling" forecast horizons.

Table 1

$$^{POINT}FP^{j}_{t,t+h} = \alpha_j + \lambda_j\left(\overline{FP_{t,t+h}}\right) + \varepsilon^{j}_{t,t+h}$$

Point Forecast Data	Equal Predictive Ability $H_0 : \alpha_j = 0 \text{ and } \lambda_j = 1$	Normalization Approach $H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_N = 0$	Time Fixed Effects $H_0 : \lambda_1 = \lambda_2 = \ldots = \lambda_N$
GDP growth: one-year-ahead	$\chi^2(62) = 375.4$ **	$\chi^2(31) = 84.4$ **	$\chi^2(30) = 278.8$ **
GDP growth: one-year/one-year-forward	$\chi^2(58) = 244.7$ **	$\chi^2(29) = 68.1$ **	$\chi^2(28) = 143.9$ **
Inflation: one-year-ahead	$\chi^2(68) = 77.7$	$\chi^2(34) = 41.1$	$\chi^2(33) = 29.3$
Inflation: one-year/one-year-forward	$\chi^2(62) = 156.2$ **	$\chi^2(31) = 90.9$ **	$\chi^2(30) = 80.4$ **
Unemployment: one-year-ahead	$\chi^2(56) = 134.5$ **	$\chi^2(28) = 58.0$ **	$\chi^2(27) = 58.1$ **
Unemployment: one-year/one-year-forward	$\chi^2(48) = 225.8$ **	$\chi^2(24) = 42.0$ *	$\chi^2(23) = 100.1$ **

Note: Degrees of freedom are reported in parentheses.
** Significant at the 1% level
* Significant at the 5% level

Table 2

$$^{DENSITY}FP^{j}_{t,t+h} = \alpha_j + \lambda_j\left(\overline{FP_{t,t+h}}\right) + \varepsilon^{j}_{t,t+h}$$

Density Forecast Data	Equal Predictive Ability $H_0 : \alpha_j = 0 \text{ and } \lambda_j = 1$	Normalization Approach $H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_N = 0$	Time Fixed Effects $H_0 : \lambda_1 = \lambda_2 = \ldots = \lambda_N$
GDP growth: one-year-ahead	$\chi^2(62) = 337.5^{**}$	$\chi^2(31) = 235.5^{**}$	$\chi^2(30) = 160.6^{**}$
GDP growth: one-year/one-year-forward	$\chi^2(58) = 371.2^{**}$	$\chi^2(29) = 173.1^{**}$	$\chi^2(28) = 100.6^{**}$
Inflation: one-year-ahead	$\chi^2(68) = 352.6^{**}$	$\chi^2(34) = 208.8^{**}$	$\chi^2(33) = 103.2^{**}$
Inflation: one-year/one-year-forward	$\chi^2(62) = 345.6^{**}$	$\chi^2(31) = 273.5^{**}$	$\chi^2(30) = 134.3^{**}$
Unemployment: one-year-ahead	$\chi^2(56) = 324.2^{**}$	$\chi^2(28) = 141.1^{**}$	$\chi^2(27) = 106.0^{**}$
Unemployment: one-year/one-year-forward	$\chi^2(48) = 138.6^{**}$	$\chi^2(24) = 69.8^{**}$	$\chi^2(23) = 74.6^{**}$

Note: Degrees of freedom are reported in parentheses.
** Significant at the 1% level
* Significant at the 5% level

Table 3

Rank Orderings of Forecast Accuracy: Point Forecasts of One-Year-Ahead Unemployment Rate

Forecaster ID	$\overline{FP} = 0.25$	$\overline{FP} = 0.50$	$\overline{FP} = 0.75$	$\overline{FP} = 1.25$	$\overline{FP} = 2.0$
1	1	5	15	20	23
37	2	2	5	9	12
96	3	1	2	6	7
2	4	16	23	27	27
89	5	7	12	17	18
23	6	6	11	14	16
26	7	14	20	21	22
54	8	17	22	23	25
20	9	4	8	8	8
4	10	15	17	18	20
95	11	12	10	10	9
94	12	8	7	7	6
16	13	13	13	11	10
39	14	3	4	3	3
38	15	19	18	15	15
24	16	10	6	4	5
5	17	22	19	19	17
15	18	18	14	12	11
98	19	21	16	13	14
56	20	23	24	25	24
33	21	24	26	26	26
31	22	25	25	24	21
22	23	11	3	2	2
52	24	20	9	5	4
91	25	9	1	1	1
42	26	26	21	16	13
36	27	28	28	28	28
29	28	27	27	22	19

Figure 1

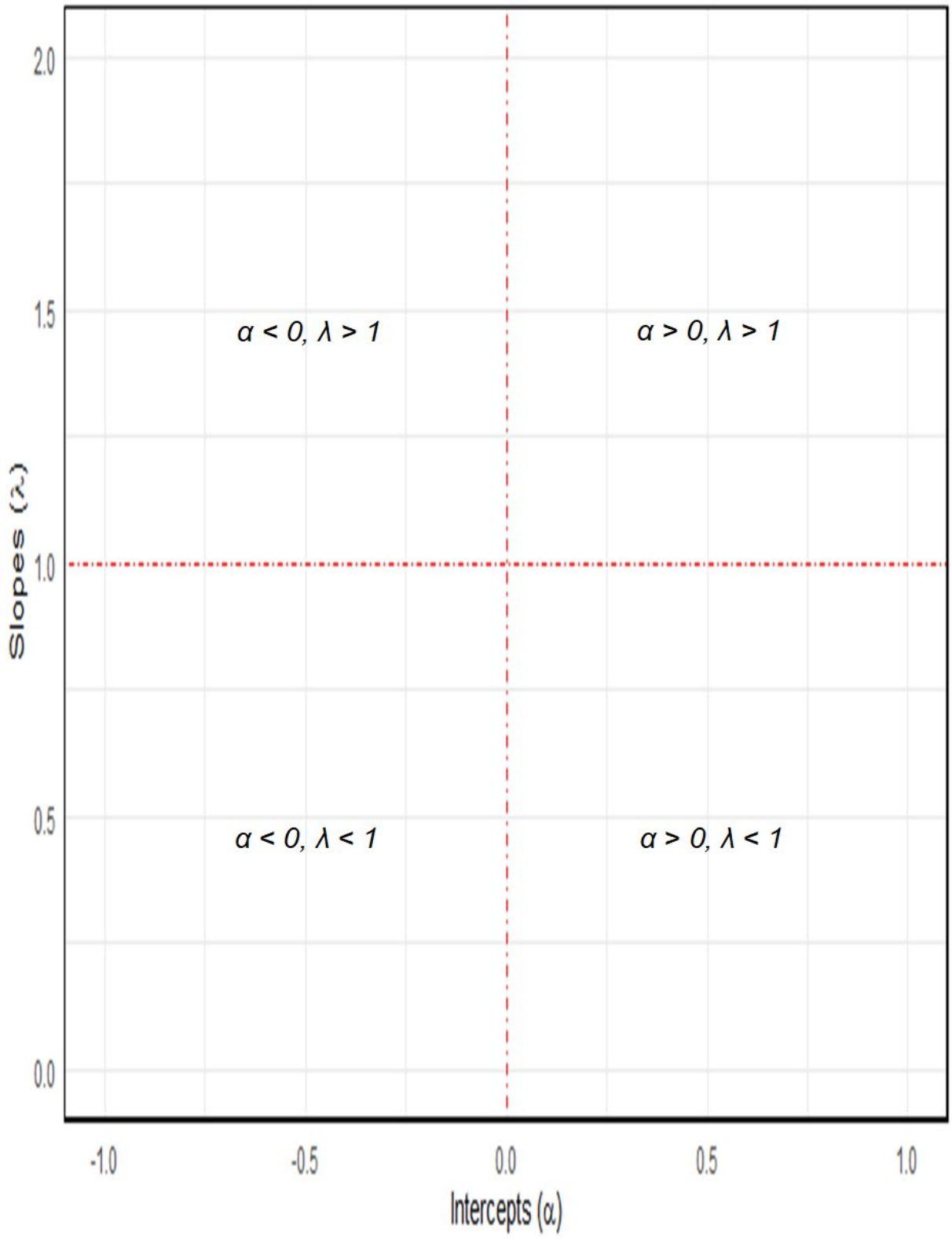

2.0
1.5
1.0
0.5
0.0
Slopes (λ)
α < 0, λ > 1
α > 0, λ > 1
α < 0, λ < 1
α > 0, λ < 1
-1.0
-0.5
0.0
0.5
1.0
Intercepts (α)

Figure 2

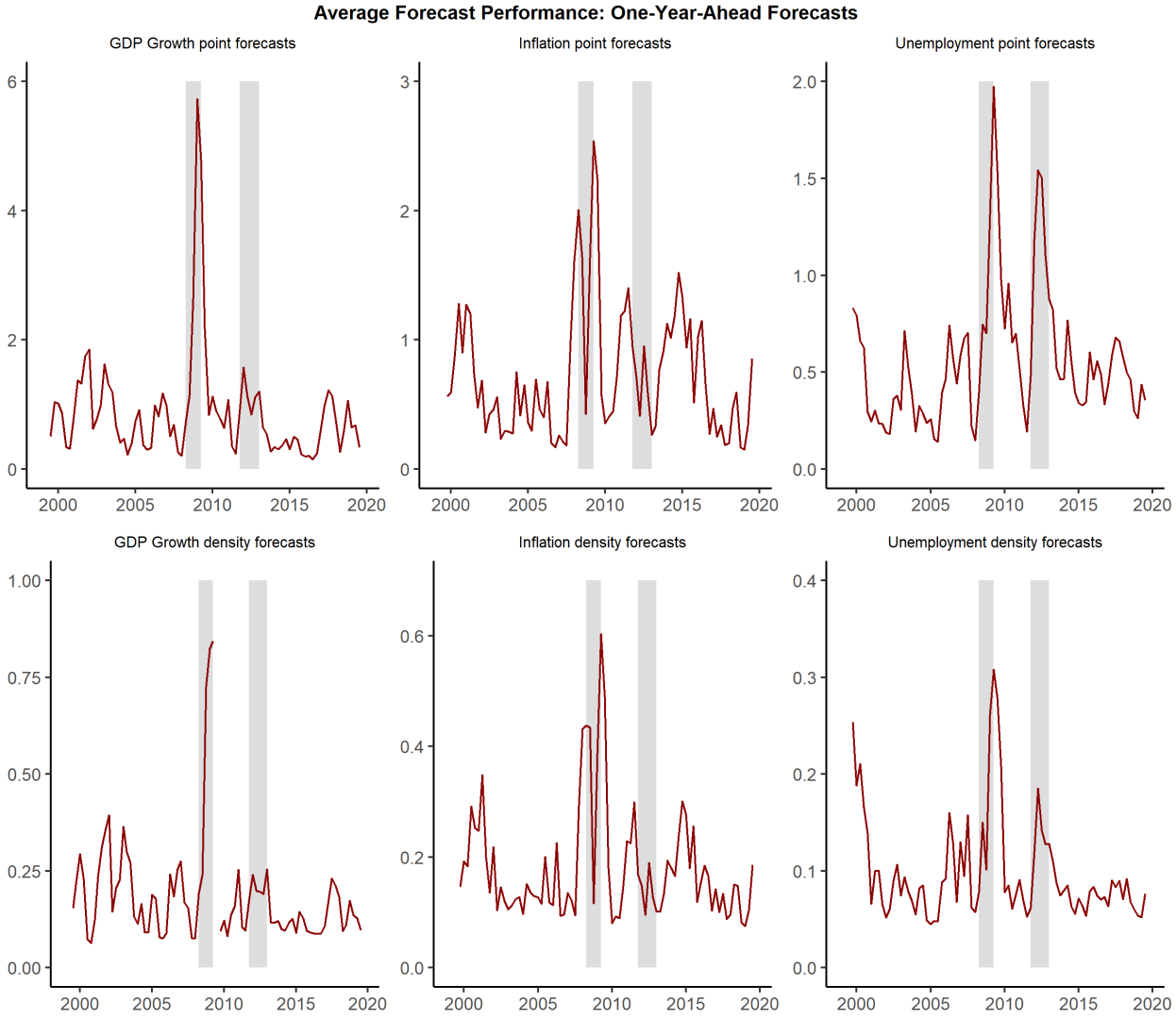

Average Forecast Performance: One-Year-Ahead Forecasts
GDP Growth point forecasts
Inflation point forecasts
Unemployment point forecasts
GDP Growth density forecasts
Inflation density forecasts
Unemployment density forecasts
2000
2005
2010
2015
2020

Figure 3

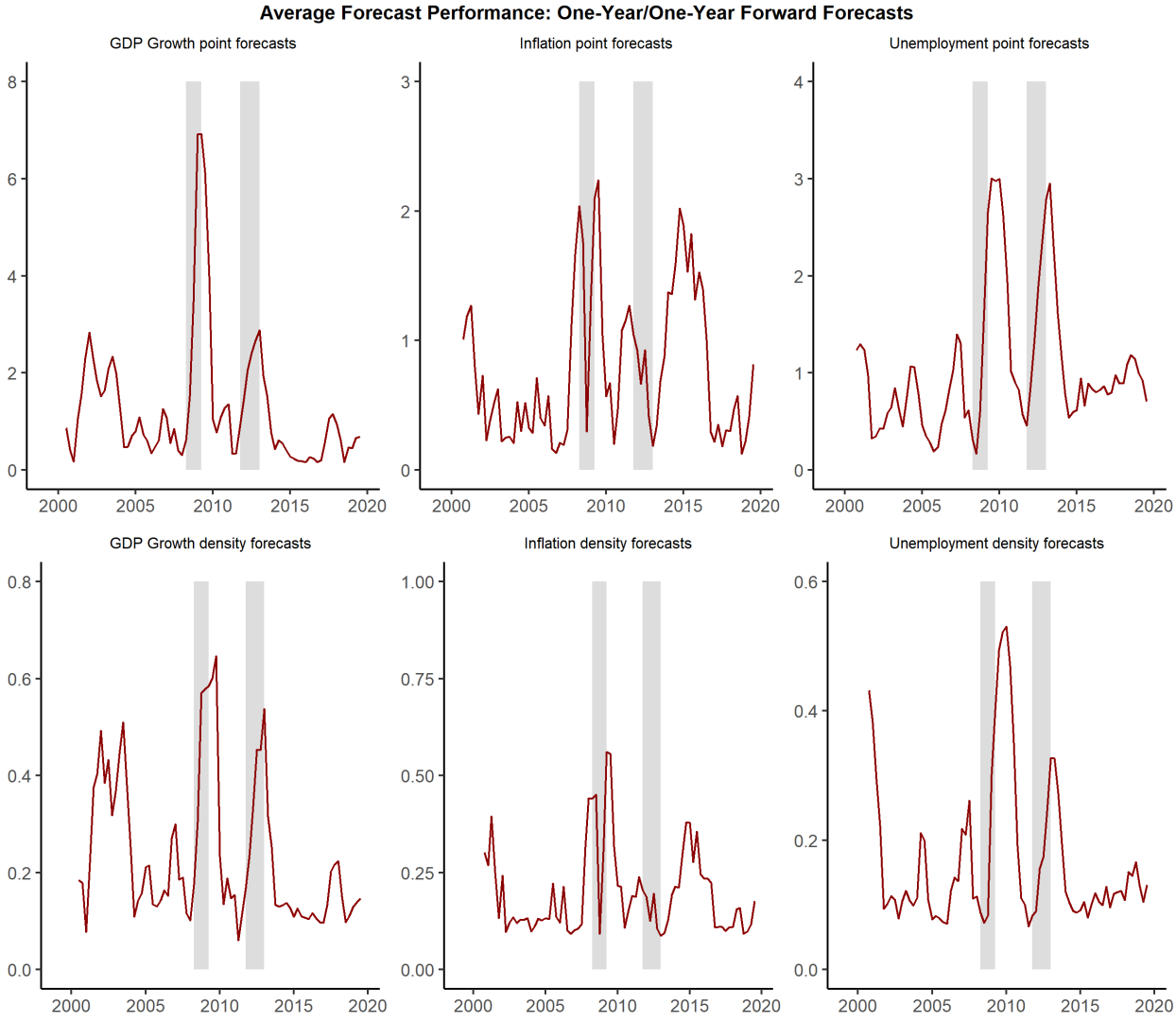

Average Forecast Performance: One-Year/One-Year Forward Forecasts
GDP Growth point forecasts
Inflation point forecasts
Unemployment point forecasts
GDP Growth density forecasts
Inflation density forecasts
Unemployment density forecasts
2000
2005
2010
2015
2020

Figure 4

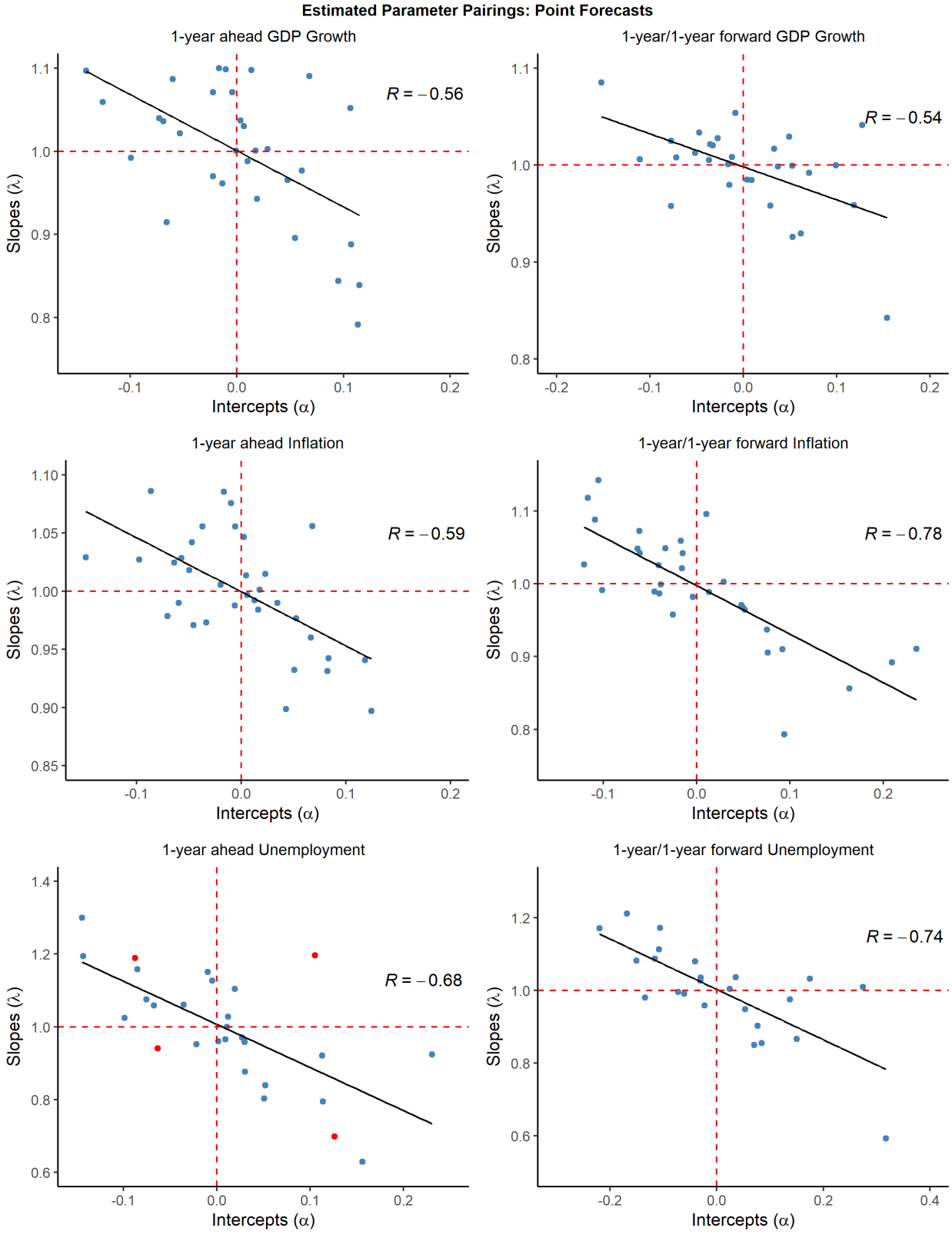

Estimated Parameter Pairings: Point Forecasts
1-year ahead GDP Growth
R = −0.56
1-year/1-year forward GDP Growth
R = −0.54
1-year ahead Inflation
R = −0.59
1-year/1-year forward Inflation
R = −0.78
1-year ahead Unemployment
R = −0.68
1-year/1-year forward Unemployment
R = −0.74
Slopes (λ)
Intercepts (α)

Figure 5

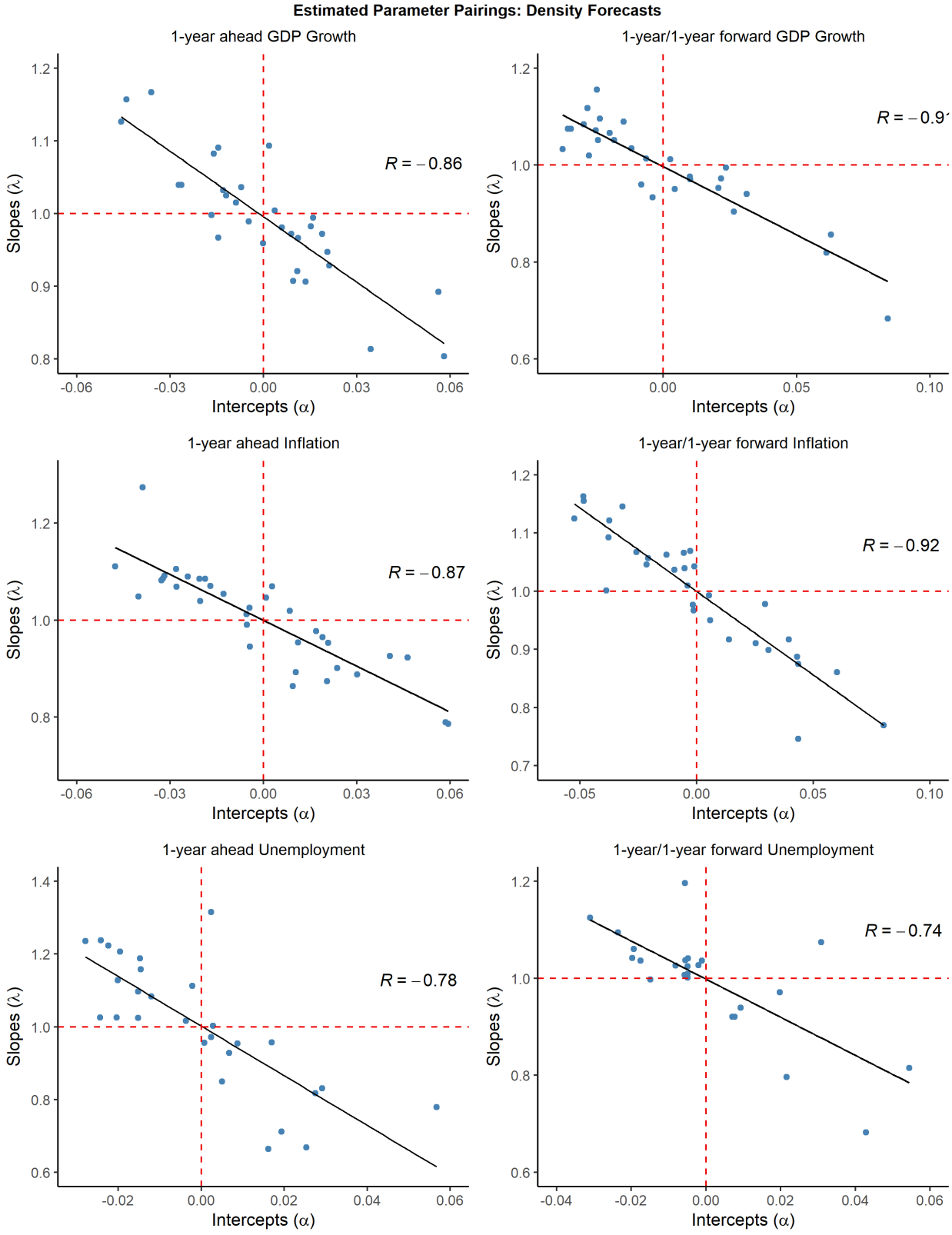
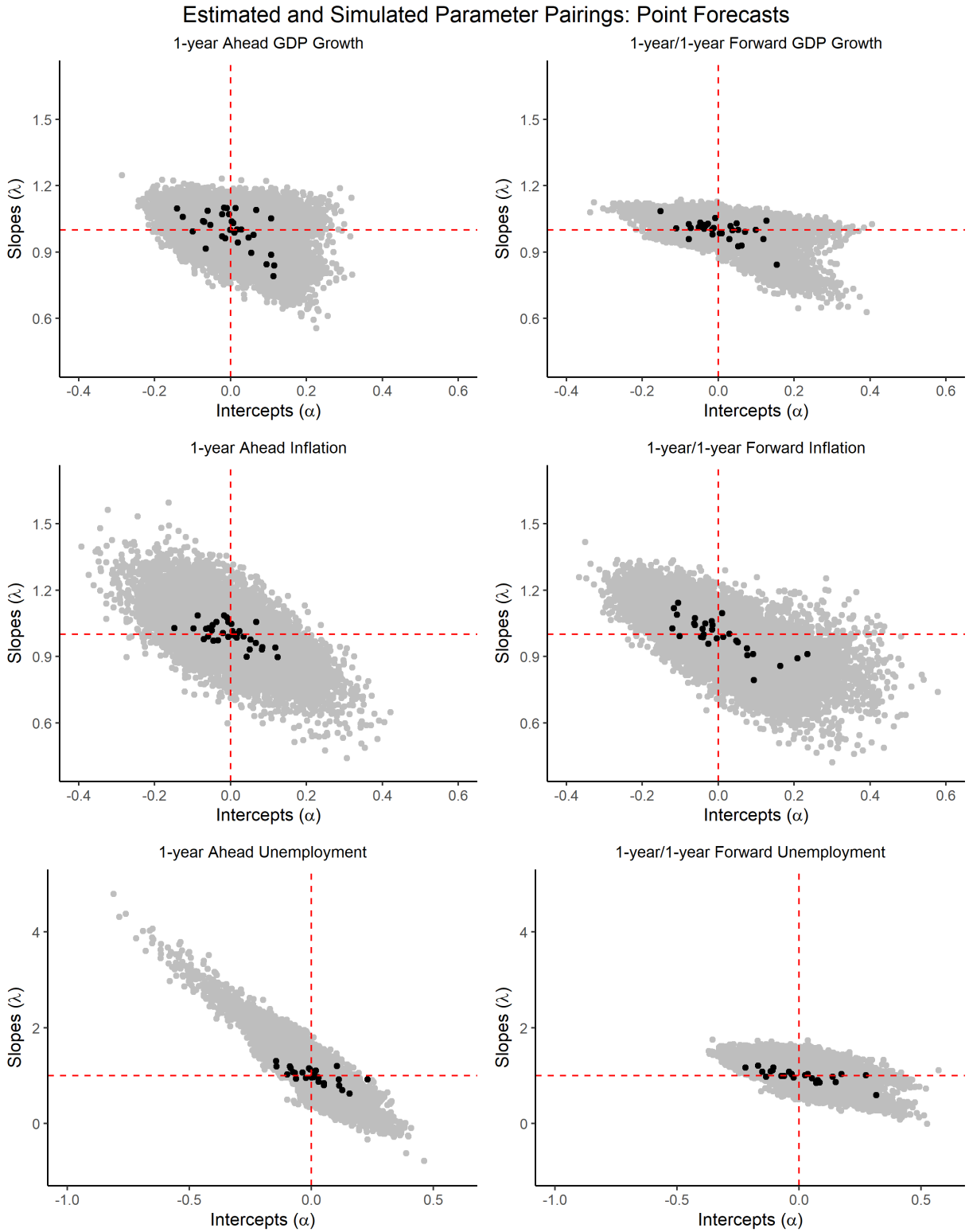
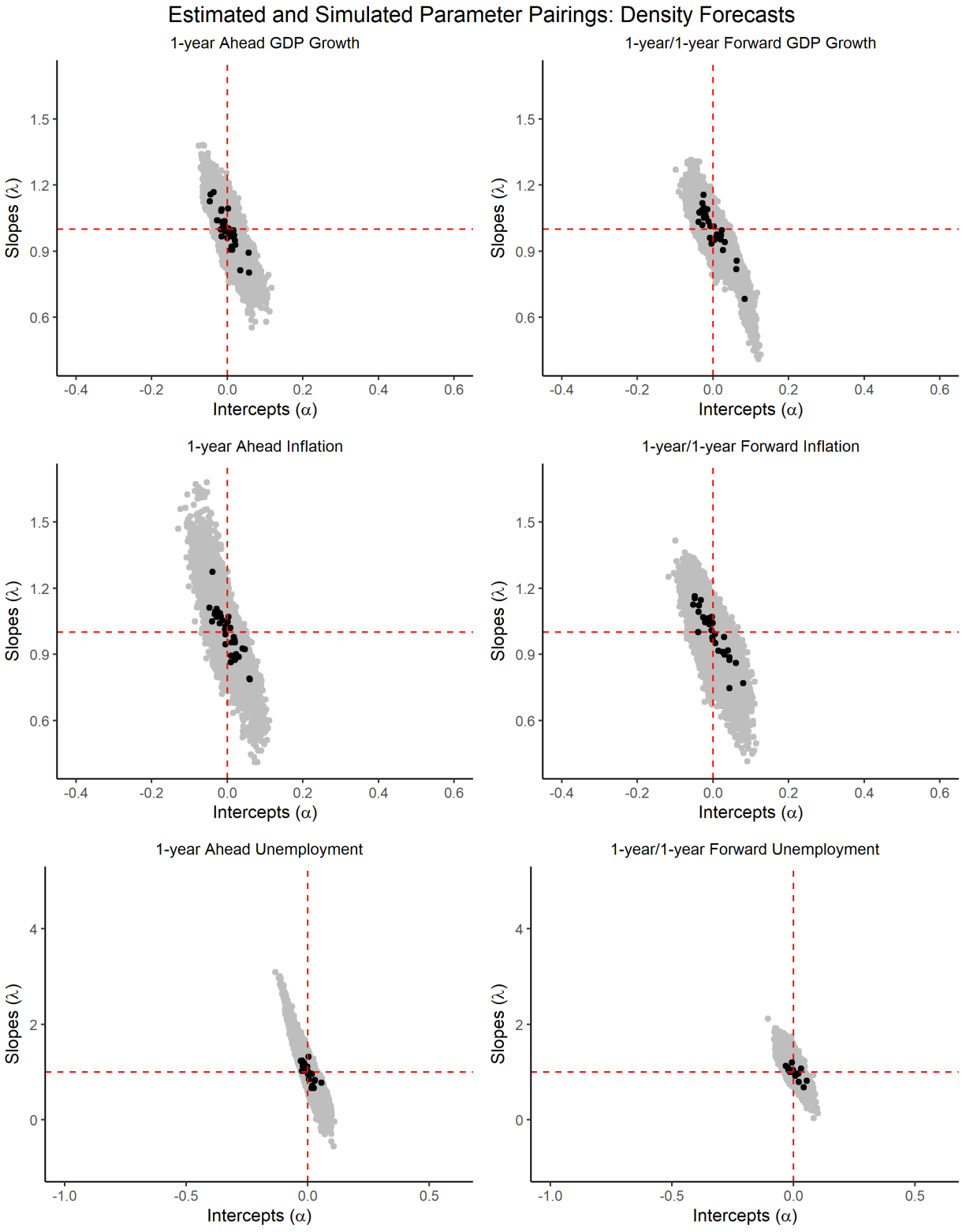

Estimated Parameter Pairings: Density Forecasts
1-year ahead GDP Growth
R = −0.86
1-year/1-year forward GDP Growth
R = −0.9
1-year ahead Inflation
R = −0.87
1-year/1-year forward Inflation
R = −0.92
1-year ahead Unemployment
R = −0.78
1-year/1-year forward Unemployment
R = −0.74
Slopes (λ)
Intercepts (α)

Figure 6



Note: Black dots are estimated values and grey dots are simulated values from the estimated joint distributions.

Figure 7



Note: Black dots are estimated values and grey dots are simulated values from the estimated joint distributions.

Figure 8

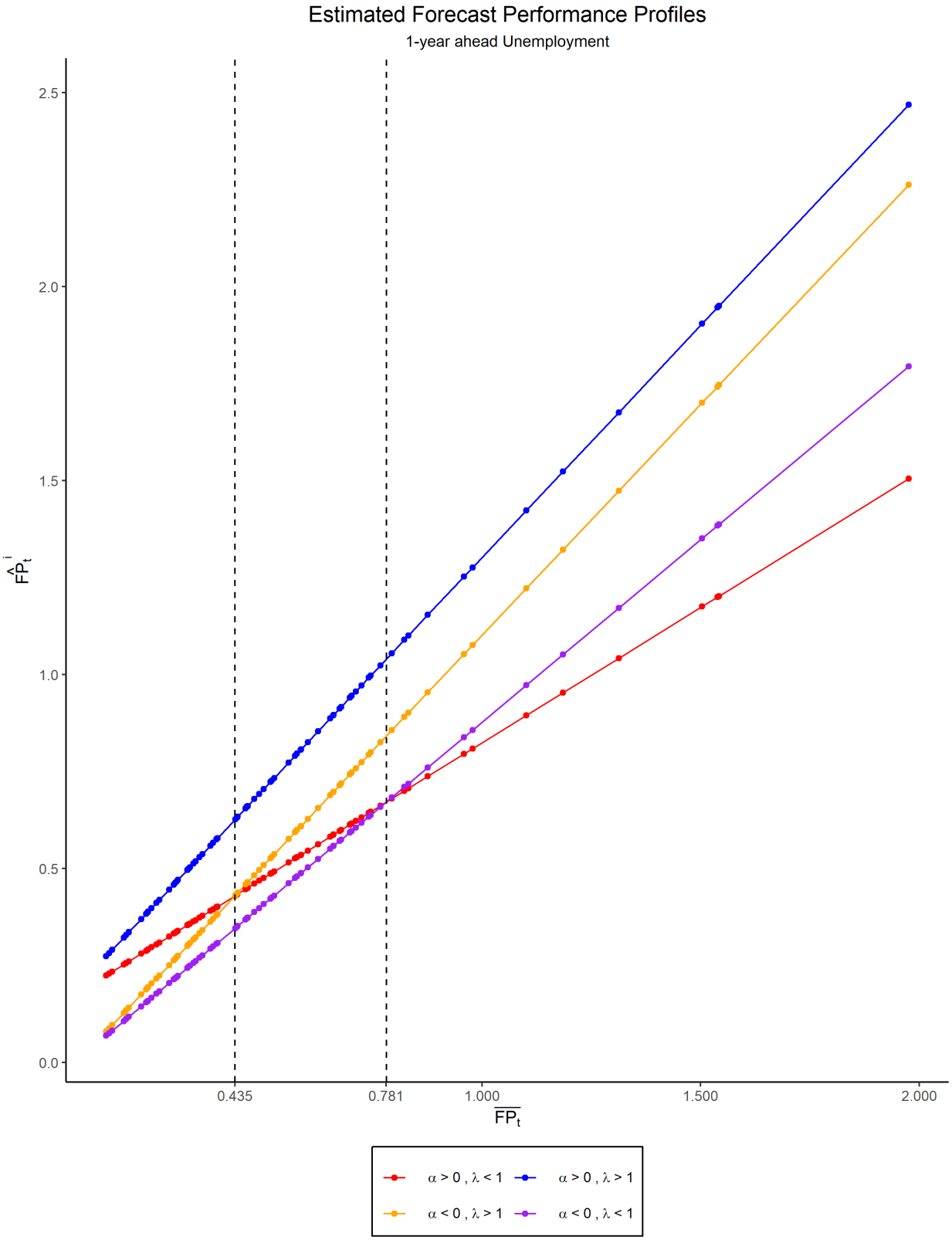

Estimated Forecast Performance Profiles
1-year ahead Unemployment
2.5
2.0
1.5
1.0
0.5
0.0
$\hat{FP}_t^i$
0.435
0.781
1.000
1.500
2.000
$\overline{FP}_t$
α > 0 , λ < 1
α > 0 , λ > 1
α < 0 , λ > 1
α < 0 , λ < 1

Figure 9

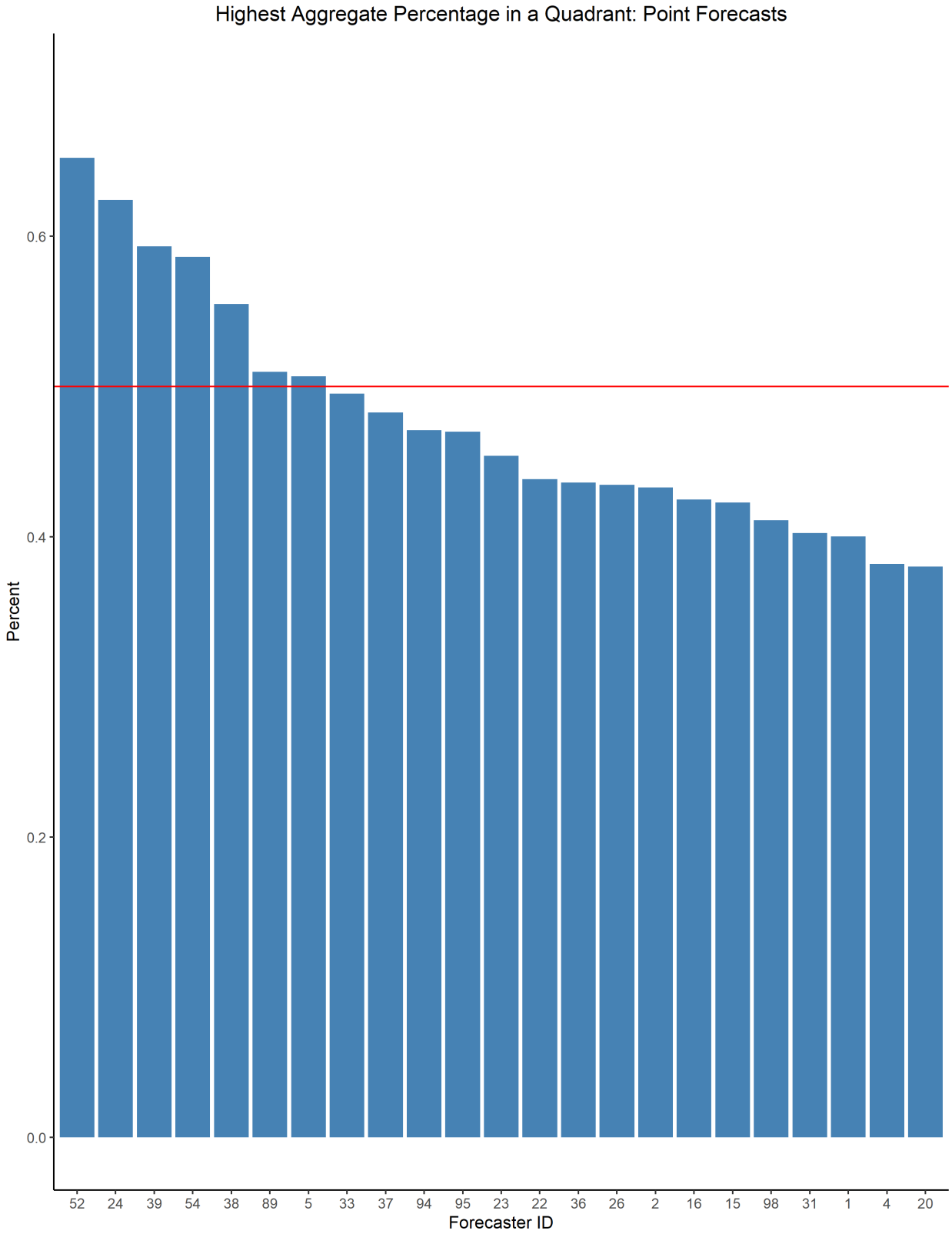

Highest Aggregate Percentage in a Quadrant: Point Forecasts
Percent
0.0
0.2
0.4
0.6
52
24
39
54
38
89
5
33
37
94
95
23
22
36
26
2
16
15
98
31
1
4
20
Forecaster ID

Figure 10

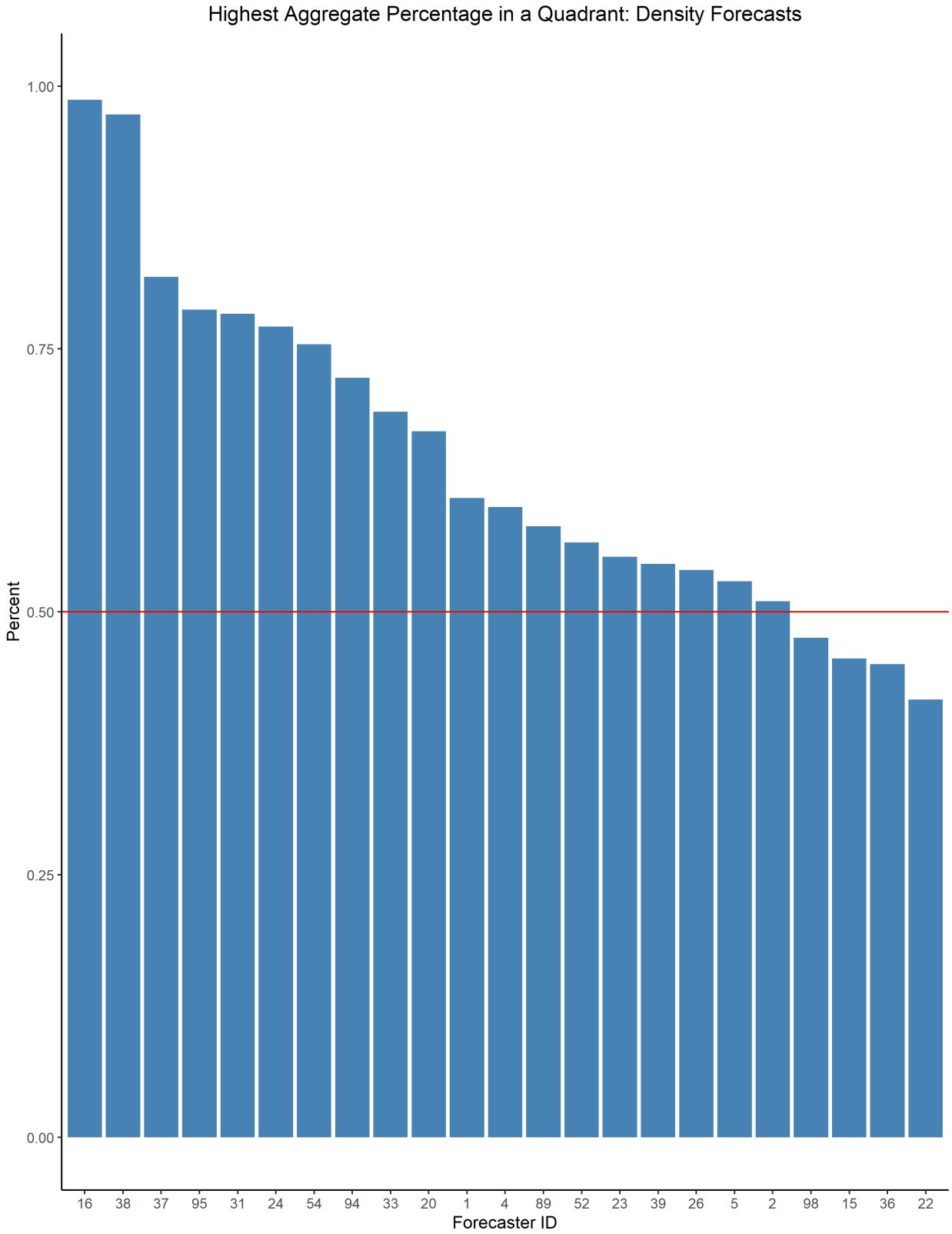
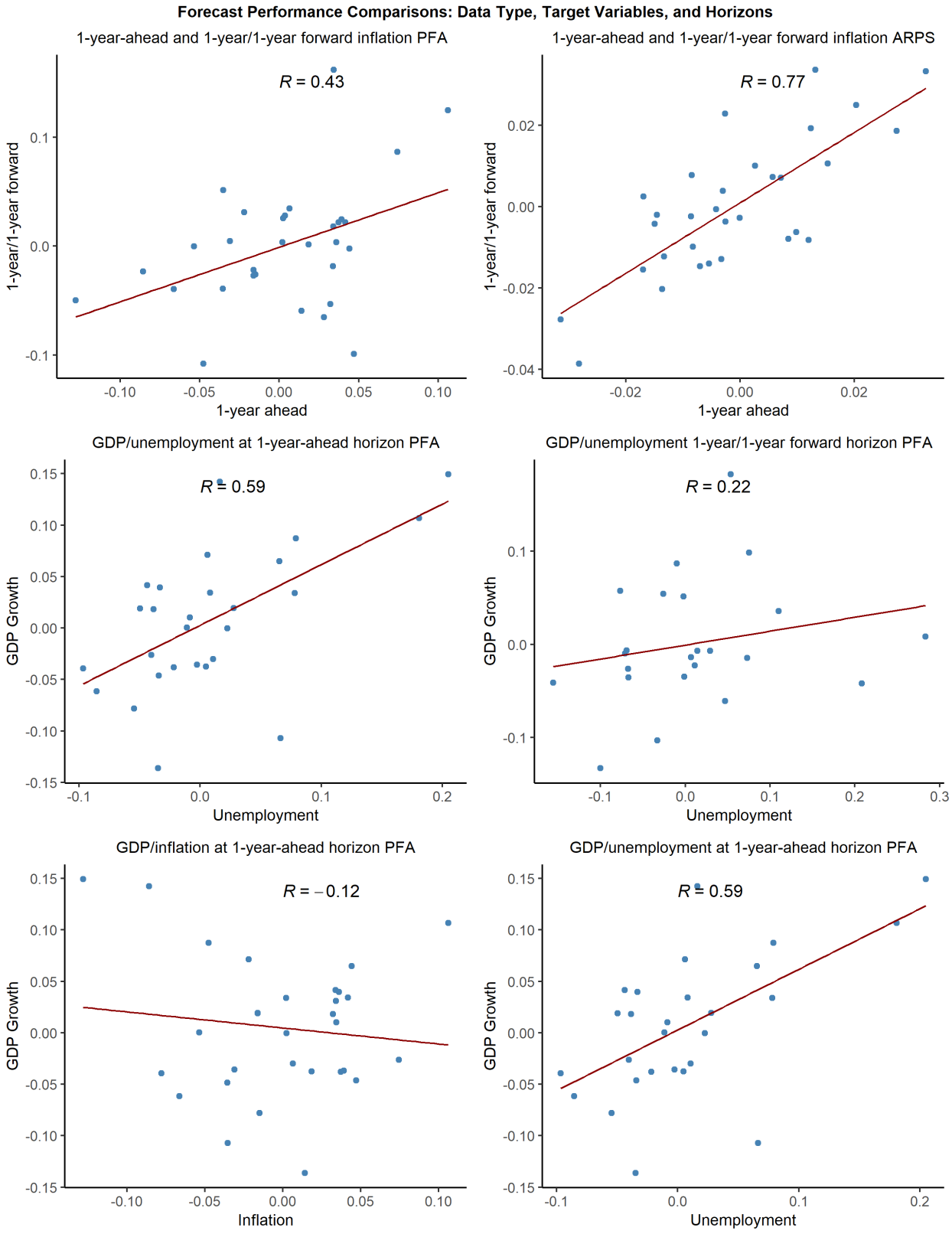

Highest Aggregate Percentage in a Quadrant: Density Forecasts
Percent
1.00
0.75
0.50
0.25
0.00
16 38 37 95 31 24 54 94 33 20 1 4 89 52 23 39 26 5 2 98 15 36 22
Forecaster ID

Figure 11


Forecast Performance Comparisons: Data Type, Target Variables, and Horizons
1-year-ahead and 1-year/1-year forward inflation PFA
R = 0.43
1-year/1-year forward
1-year ahead
1-year-ahead and 1-year/1-year forward inflation ARPS
R = 0.77
1-year/1-year forward
1-year ahead
GDP/unemployment at 1-year-ahead horizon PFA
R = 0.59
GDP Growth
Unemployment
GDP/unemployment 1-year/1-year forward horizon PFA
R = 0.22
GDP Growth
Unemployment
GDP/inflation at 1-year-ahead horizon PFA
R = −0.12
GDP Growth
Inflation
GDP/unemployment at 1-year-ahead horizon PFA
R = 0.59
GDP Growth
Unemployment

References

Abel, Joshua, Robert W. Rich, Joseph Song, and Joseph S. Tracy. 2016. "The Measurement and Behavior of Uncertainty: Evidence from the ECB Survey of Professional Forecasters." *Journal of Applied Econometrics* 31 (3): 533–50. https://doi.org/10.1002/jae.2430.

Batchelor, Roy A. 1990. "All Forecasters Are Equal." *Journal of Business & Economic Statistics* 8 (1): 143–44. https://doi.org/10.1080/07350015.1990.10509784.

Boero, Gianna, Jeremy Smith, and Kenneth F. Wallis. 2015. "The Measurement and Characteristics of Professional Forecasters' Uncertainty." *Journal of Applied Econometrics* 30 (7): 1029–46. https://doi.org/10.1002/jae.2400.

Bowles, Carlos, Roberta Friz, Veronique Genre, Geoff Kenny, Aidan Meyler, and Tuomas Rautanen. 2007. "The ECB Survey of Professional Forecasters (SPF) – A Review after Eight Years' Experience." Occasional Paper 59. European Central Bank. http://hdl.handle.net/10419/154512.

Bruine de Bruin, Wändi, Charles F. Manski, Giorgio Topa, and Wilbert van der Klaauw. 2011. "Measuring Consumer Uncertainty about Future Inflation." *Journal of Applied Econometrics* 26 (3): 454–78. https://doi.org/10.1002/jae.1239.

Christensen, Jens H., Francis X. Diebold, Georg H. Strasser, and Glenn D. Rudebusch. 2008. "Multivariate Comparison of Predictive Accuracy." Working paper. http://www.econ.uconn.edu/Seminar%20Series/strasser08.pdf.

Coibion, Olivier, and Yuriy Gorodnichenko. 2012. "What Can Survey Forecasts Tell Us about Information Rigidities?" *Journal of Political Economy* 120 (1): 116–59. https://doi.org/10.1086/665662.

———. 2015. "Information Rigidity and the Expectations Formation Process: A Simple Framework and New Facts." *American Economic Review* 105 (8): 2644–78. https://doi.org/10.1257/aer.20110306.

D'Agostino, Antonello, Kieran Mcquinn, and Karl Whelan. 2012. "Are Some Forecasters Really Better Than Others?" *Journal of Money, Credit and Banking* 44 (4): 715–32. https://doi.org/10.1111/j.1538-4616.2012.00507.x.

Diebold, Francis X., and Robert S. Mariano. 2002. "Comparing Predictive Accuracy." *Journal of Business & Economic Statistics* 20 (1): 134–44. https://doi.org/10.1198/073500102753410444.

Garcia, Juan Angel. 2003. "An Introduction to the ECB's Survey of Professional Forecasters." Research Report 8. ECB Occasional Paper. http://hdl.handle.net/10419/154461.

Mackowiak, Bartosz, and Mirko Wiederholt. 2009. "Optimal Sticky Prices under Rational Inattention." *American Economic Review* 99 (3): 769–803. https://doi.org/10.1257/aer.99.3.769.

Mankiw, N. Gregory, and Ricardo Reis. 2002. "Sticky Information versus Sticky Prices: A Proposal to Replace the New Keynesian Phillips Curve." *Quarterly Journal of Economics* 117 (4): 1295–1328. https://doi.org/10.1162/003355302320935034.

Mankiw, N. Gregory, Ricardo Reis, and Justin Wolfers. 2004. "Disagreement about Inflation Expectations." *NBER Macroeconomics Annual* 18: 209–48. https://doi.org/10.1086/ma.18.3585256.

Meyler, Aidan. 2020. "Forecast Performance in the ECB SPF: Ability or Chance?" 2371. Working Paper Series. European Central Bank. https://ideas.repec.org/p/ecb/ecbwps/20202371.html.

Newey, Whitney K., and Kenneth D. West. 1987. "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix." *Econometrica* 55 (3): 703–8. https://doi.org/10.2307/1913610.

Pesaran, M. Hashem. 2006. "Estimation and Inference in Large Heterogeneous Panels with a Multifactor Error Structure." *Econometrica* 74 (4): 967–1012. https://doi.org/10.1111/j.1468-0262.2006.00692.x.

Rich, Robert W., and Joseph S. Tracy. 2020. "A Closer Look at the Behavior of Uncertainty and Disagreement: Micro Evidence from the Euro Area." *Journal of Money, Credit and Banking*, September, jmcb.12728. https://doi.org/10.1111/jmcb.12728.

Sarafidis, Vasilis, and Tom Wansbeek. 2012. "Cross-Sectional Dependence in Panel Data Analysis." *Econometric Reviews* 31 (5): 483–531. https://doi.org/10.1080/07474938.2011.611458.

Sims, Christopher A. 2003. "Implications of Rational Inattention." *Journal of Monetary Economics* 50 (3): 665–90. https://doi.org/10.1016/S0304-3932(03)00029-1.

Stekler, Herman O. 1987. "Who Forecasts Better?" *Journal of Business & Economic Statistics* 5 (1): 155–58. https://doi.org/10.1080/07350015.1987.10509571.

Woodford, Michael. 2002. "Imperfect Common Knowledge and the Effects of Monetary Policy." In *Knowledge, Information, and Expectations in Modern Macroeconomics: In Honor of Edmund S. Phelps*, edited by Philippe Aghion, Roman Frydman, Joseph E. Stiglitz, and Michael Woodford. Princeton University Press.