Predicting bond return predictability

Abstract

This paper provides empirical evidence on predictable shifts in the degree of bond return predictability. Bond returns are predictable in high (low) economic activity (uncertainty) states, which suggests that the expectations hypothesis of the term structure holds periodically. These state-dependencies in predictability, established by introducing a new multivariate test for equal conditional predictive ability, can be used in real-time to improve out-of-sample bond risk premia estimates and investors' economic utility through a novel dynamic forecast combination scheme. Dynamically combined forecasts exhibit strong countercyclical behavior and peak during recessions. The empirical findings can be explained within a non-linear term structure model.

Keywords: Bond excess returns, forecasting, state-dependencies, multivariate test, equal conditional predictive ability

JEL Classification: C12, C52, E43, E44, G12

This version: September 21, 2020

I. Introduction

The expectations hypothesis (EH) of the term structure holds periodically. We establish this fact with new empirical evidence of periodic and predictable shifts in bond return predictability related to economic activity and uncertainty. Specifically, our findings suggest that bond return predictability is associated with periods of high (low) economic activity (uncertainty). These predictable shifts can be used in real-time to improve out-of-sample risk premia estimates and investors' economic utility. Existing evidence on bond return predictability has traditionally been established using linear predictive regressions designed to assess whether bond returns are predictable *on average* using time series that potentially span many diverse states of nature.[1] If predictability shifts over time, then an unconditional approach may be misleading and lead to distorted inference. For example, while in-sample evidence frequently points to predictability by means of variables such as forward spreads (Fama and Bliss, 1987), yield spreads (Campbell and Shiller, 1991), forward rates (Cochrane and Piazzesi, 2005), and macroeconomic variables (Cooper and Priestley, 2009, Ludvigson and Ng, 2009, Cieslak and Povala, 2015, Eriksen, 2017), out-of-sample exercises often fail to deliver consistent evidence of predictability and statistical and economic evaluations often disagree (Thornton and Valente, 2012, Sarno, Schneider, and Wagner, 2016). Our empirical findings suggest that accounting for state-dependencies in bond return predictability can resolve these puzzling contradictions.

We address the issue of time-varying bond return predictability by adopting a *conditional* perspective on predictability and using observable state variables to identify methods anticipated to be informative of future relative forecast performance in a new modeling

---

[1]Early studies include Fama and Bliss (1987), Keim and Stambaugh (1986), Fama and French (1989), and Campbell and Shiller (1991). More recent studies of bond return predictability includes Cochrane and Piazzesi (2005), Cooper and Priestley (2009), Ludvigson and Ng (2009), Cieslak and Povala (2015), Eriksen (2017), Ghysels, Horan, and Moench (2018), Gargano, Pettenuzzo, and Timmermann (2019), Berardi, Markovich, Plazzi, and Tamoni (2020), and Bianchi, Büchner, and Tamoni (2020).

framework.[2] Our contributions are fivefold. First, we provide new empirical evidence on predictable state-dependencies in bond return predictability. We document that bond return predictability shifts over time for a collection of standard bond predictors from the literature including yield curve information and macroeconomic variables. We begin with a traditional evaluation of out-of-sample forecasts using *unconditional* predictive ability tests and find little evidence that individual predictors are able to reliably outperform the EH. Importantly, however, this does not exclude the possibility that a given method may display superior predictability in certain states of the world. To facilitate a conditional, state-dependent view of bond return predictability, we introduce a multivariate generalization of the statistical test for equal (un)conditional predictive ability among two or more forecasting methods presented in Giacomini and White (2006).[3] We employ the test to assess differences in *conditional* predictive abilities and find overwhelming evidence favoring state-dependencies in bond return predictability.

Second, we show that the shifts in predictability are related to economic activity and uncertainty measured using the Purchasing Manager's Index (PMI) (see, e.g., Berge and Jordà (2011) and Christiansen, Eriksen, and Møller (2014)) and the macroeconomic uncertainty index ($\mathcal{U}$) proposed in Jurado, Ludvigson, and Ng (2015), respectively.[4] We uncover a striking pattern in bond return predictability across states related to these variables. Interpreting the EH as a no-predictability benchmark, which is standard practice in the literature, we

---

[2]Predictability of future relative forecast performance is even more relevant when viewed in the light of the numerous studies that provide empirical evidence of model instabilities in predictive models. Prominent examples include Pesaran, Pettenuzzo, and Timmermann (2006), Giacomini and Rossi (2009, 2010), Pettenuzzo and Timmermann (2011), Rossi (2013), and Pettenuzzo and Timmermann (2017).

[3]The test also extends the (unconditional) multivariate Diebold and Mariano (1995) statistic proposed in Mariano and Preve (2012) by allowing for comparison of a mixture of nested and non-nested models.

[4]A number of alternative measures of uncertainty have been proposed. The qualitative findings presented in this paper remain unchanged if the Economic Political Uncertainty index by Baker, Bloom, and Davis (2016) or the VXO index by CBOE are used in place of the index by Jurado et al. (2015).

provide evidence that bond risk premia are predictable in high (low) economic activity (uncertainty) states. Conversely, time invariant risk premia, as implied by the EH, provides a reasonable approximation in low (high) economic activity (uncertainty) states. Consistent with this, out-of-sample R^2s (Campbell and Thompson, 2008) for individual predictors are mostly negative in low (high) economic activity (uncertainty) states and positive in high (low) activity (uncertainty) states. In short, albeit several predictors fail to provide valuable information *on average*, many outperform the EH *conditional* on the state of the economy.

Third, we show that the predictable state-dependencies in bond return predictability are exploitable for real-time forecasting purposes. We propose a simple dynamic ranking rule for identifying the set of forecasting methods that possess the best, yet statistically indistinguishable, conditional predictive ability in real-time inspired by the Model Confidence Set (MCS) approach of Hansen, Lunde, and Nason (2011). As we shall see, our multivariate test for conditional predictability ability is instrumental in identifying and eliminating inferior methods by informing us of state-dependent differences. The rule is based on least squares predictions of relative forecasting performance and is straightforward to implement. Importantly, our method is based on expected predictive ability, which separates it from methods relying on past performance (see, e.g., Aiolfi and Timmermann (2006), Samuels and Sekkel (2017), and Adämmer and Schüssler (2020)). We show that this is a key feature in correctly identifying conditional predictability in our setting. The best set of methods at any given point in time may then consists of a single method, all methods, or any number in between. If more than one method is selected, we perform equal-weighted forecast combination (Bates and Granger, 1969) among the selected methods. It is well established that an equal-weighted combination strategy among all initial methods is hard to beat (Timmermann, 2006) and performs well for financial time series (Rapach, Strauss, and Zhou, 2010). Our approach can thus be viewed as a data-driven and forward-looking way to select between dynamic model selection and forecast combination. It further addresses an outstanding issue raised in Aiolfi, Capistrán, and Timmermann (2011) that little attention has been paid to determining

the optimal set of models to combine given a potential pool of candidate predictors. Our dynamic approach provides a natural way to select the methods expected to perform well and dynamically eliminate those expected to perform poorly.[5] We show that the implementation of the forecast combination scheme with dynamic trimming delivers sizeable gains in predictive accuracy relative to (i) the EH, (ii) a static equal-weighted forecast combination strategy, and (iii) a dynamic combination strategy based on past (unconditional) average predictability when evaluated using both standard statistical criteria and when measuring the economic value from the viewpoint of a mean-variance investor that trades in the Treasury bond market. Importantly, dynamically combined forecasts deliver positive and sizeable out-of-sample R^2 values and utility gains even in states where all individual predictors are weak. In fact, dynamically combined forecasts always performs better than the EH in all states and at least as good as the competing combination strategies. This suggests that a conditional view of bond return predictability offers a resolution to the puzzling contradiction between the statistical and economic evidence often found in the literature (Thornton and Valente, 2012, Sarno et al., 2016). Importantly, this is achieved within a standard linear framework and without having to adopt Bayesian approaches with stochastic volatility (Gargano et al., 2019) or computationally expensive machine learning methods (Bianchi et al., 2020).

Fourth, we document that our dynamic forecast combination scheme generates out-of-sample bond risk premia estimates that are strongly countercyclical and peak in recessions. This is an important result as nearly all individual predictors (except the Ludvigson and Ng (2009) factor) generates procyclical risk premia estimates. Unlike the individual predictors, the dynamically combined forecasts are therefore consistent with standard finance theory in which risk premia to expected be high in bad economic times due to heightened risk aversion

---

[5]A large empirical literature documents gains from (statically) trimming the sets of forecast methods prior to averaging. Notable examples include Aiolfi and Favero (2005), Aiolfi and Timmermann (2006), Timmermann (2006), Stock and Watson (2004), Rapach et al. (2010), Bjørnland, Gerdrup, Jore, Smith, and Thorsrud (2012), and Genre, Kenny, Meyler, and Timmermann (2013).

(Campbell and Cochrane, 1999, Wachter, 2006, Buraschi and Jiltsov, 2007), and support models with time-varying risk and risk prices (Bekaert, Engstrom, and Xing, 2009, Carpenter, Lu, and Whitelaw, 2020, Creal and Wu, 2020). The simple equal-weighted combination schemes, on the other hand, generates acyclical out-of-sample forecasts with little relation to the macroeconomy. The fact that our dynamic forecast combination scheme delivers countercyclical out-of-sample risk premia forecasts that improve overall predictive accuracy and economic value strongly supports our conclusion that the test reliably identifies and exploits shifts in bond returns predictability.

Last, we illustrate that our empirical findings can be explained within a non-linear term structure model that allows for state-dependencies in the yield curve. In particular, we consider a one-factor version of the non-linear model presented in Feldhütter, Heyerdahl-Larsen, and Illeditsch (2018) in which bond prices, short rates, and prices of risk can be viewed as linear combination of the corresponding values in two artificial economies, which we interpret as corresponding to two different states of the world. The estimated model predicts that the EH holds in one of two states, but not in the other, which is consistent with state-dependent shifts in bond return predictability. The model-implied stochastic weight on the state of the economy in which the EH holds is decreasing (increasing) with economic activity (uncertainty), which is consistent with the patterns document in this paper. The model therefore provides a simple, but clear, illustration of a mechanism that can rationalize our empirical findings. Moreover, it is possible to interpret the stochastic state weight as a measure of disagreement between agents in a reduced-form consumption-based asset pricing model. Under this view, the model allows for an assessment of time-varying differences in beliefs among agents, which Cujean and Hasler (2017) show can generate state-dependent predictability in a theoretical framework.

In sum, we provide new empirical evidence of predictable state-dependencies in bond return predictability that are linked to economic activity and uncertainty. We document that these predictability shifts are exploitable in real-time and delivers sizable gains in both

predictive accuracy and economic value. The gains originate from the method's ability to correctly predict relative forecasting performance and this, in turn, leads to better and economically meaningful out-of-sample bond risk premia estimates.

Related literature Our paper relates to several strands of literature. First, an extensive literature studies the predictability of Treasury bond excess returns. This literature has traditionally evaluated predictability from an *unconditional* perspective. We contribute to this literature by offering a *conditional* perspective on bond return predictability. Our paper is therefore closely related to a large and active literature on time-varying predictability of asset returns. Rapach et al. (2010), Henkel, Martin, and Nardari (2011), and Dangl and Halling (2012) find that stock return predictability is closely linked to the business cycle and mostly present in recessions. Cujean and Hasler (2017) offer a theoretical model for this observation based on time-varying disagreement among investors. Farmer, Schmidt, and Timmermann (2019) show that predictability is mainly local in time and consists of "pockets" with significant predictability. Gargano et al. (2019) argue that bond predictability is stronger in recessions, but present in both states. A similar result is found in Bianchi et al. (2020). Andreasen, Engsted, Møller, and Sander (2020) find in-sample evidence for time-varying parameters in bond prediction models and show that bond risk premia relate positively (negatively) to yield spreads in expansions (recessions). These findings are generally consistent with studies such as Hamilton (1988), Gray (1996), and Ang and Bekaert (2002) who provide evidence of regime switches in interest rates which are related to economic activity. We add to this literature by providing new evidence on the real-time predictability of time-variations in out-of-sample forecasting performance for a set of standard bond predictors. Our empirical evidence further contributes to our understanding of bond market dynamics by demonstrating that bond return predictability is stronger in high (low) economic activity (uncertainty) states and that exploiting this using a dynamic forecast combination strategy yields highly countercyclical out-of-sample forecasts in accordance with standard asset pricing

theory. Although our results differ from Gargano et al. (2019), a direct comparison is not possible as our results are based on out-of-sample estimates, whereas their decomposition is based on full-sample estimates. Moreover, our low (high) activity states are, like the pockets of Farmer et al. (2019), more encompassing than the recession (expansion) periods defined by the NBER. We further argue that adopting a conditional view of predictability offers a solution to the puzzling contradiction between statistical and economic evaluations of predictability (Thornton and Valente, 2012, Sarno et al., 2016).

Our paper is also related to a growing literature on forecast evaluation. Our multivariate generalization of the Giacomini and White (2006) test provides forecasters with the opportunity to test equal (un)conditional predictive ability among many forecast methods without having to rely on multiple testing adjustments (Hubrich and West, 2010) or non-standard and context-specific distribution often found in the literature (Clark and McCracken, 2001, McCracken, 2007, Clark and McCracken, 2012, Gonçalves, McCracken, and Perron, 2017).[6] Our tests are applicable to a mixture of both nested and non-nested models, hold for a general loss function, allow for non-stationarity in the data, and permit comparison of a wider class of forecasting methods than those considered in this paper, including linear, non-linear, Bayesian, and non-parametric methods.[7] Finally, we contribute to the literature on forecast combination and dynamic method selection. Early contributions include Makridakis and Winkler (1983) and more recent contributions documenting the empirical benefits include Jose and Winkler (2008), Rapach et al. (2010), Samuels and Sekkel (2017), Diebold and Shin (2019), and Adämmer and Schüssler (2020). Samuels and Sekkel (2017) find that using the (unconditional) MCS as a trimming device prior to constructing combined forecasts

---

[6]Moreover, our tests are generally invariant to any reordering of the forecasting methods under comparison, ensuring that conclusions drawn from a single test is unaltered by any permutation of the ordering of the forecasting methods. This alleviates the need for multiple testing adjustments.

[7]That the test allows for nested models is especially important in our setting as the EH is nested in all of our predictive methods. This sets the test apart from those presented in Hubrich and West (2010), Mariano and Preve (2012), and Clark and McCracken (2012).

can greatly improve accuracy. Diebold and Shin (2019) propose a LASSO-based procedure that sets some combining weights to zero and shrinks the remaining weights toward equal-weighting. Adämmer and Schüssler (2020) perform model selection or forecast combination by ranking multiple methods dynamically based on their past predictability. Our approach differs by being rooted in a formal multivariate test of equal conditional predictive ability and by focusing on future expected performance (Timmermann and Zhu, 2017, Granziera and Sekhposyan, 2019) rather than past (unconditional) performance. For comparison, we implement a version of the unconditional dynamic method selection rule (Samuels and Sekkel, 2017, Adämmer and Schüssler, 2020) and find that a conditional procedure provides superior predictive ability and that, in our setting, the unconditional version does not improve upon the static equal-weighed combination rule.

The remainder of the paper proceeds as follows. Section II outlines the setting and data. Section III develops a multivariate tests for equal (un)conditional predictive ability and introduces the dynamic forecast combination scheme. Section IV presents our main empirical findings on state-dependencies in bond return predictability. Section V examines the link between out-of-sample risk premia estimates and the real economy. Section VI quantifies the economic value of predictable state-dependencies. Section VII presents a non-linear term structure model that generates state-dependent predictability. Finally, Section VIII provides concluding remarks.

II.   Bond return predictability

This section describes the predictive regression framework for bond excess returns and provides summary statistics. We then outline the bond predictors used in the empirical analyses and discuss the state variables used to assess state-dependencies in bond return predictability.

A.   Predictive regression for bond returns

We consider a classic linear predictive regression model for bond risk premia

$$rx_{t+\tau}^{(k)} = \alpha^{(k)} + \boldsymbol{\beta}^{(k)}\boldsymbol{x}_t + \varepsilon_{t+\tau}^{(k)}, \tag{1}$$

where $rx_{t+\tau}^{(k)} = p_{t+\tau}^{(k-\tau)} - p_t^{(k)} - p_t^{(\tau)}$ denotes the τ-month log excess holding period return on a k-month zero-coupon Treasury bond and $p_t^{(k)}$ is the time-t log price of a bond with k months to maturity. We are interested in determining whether a set of predictors $\boldsymbol{x}_t$ can improve the prediction of bond excess returns relative to the constant expected returns benchmark $\boldsymbol{\beta}^{(k)} = \boldsymbol{0}$ (i.e. no predictability) implied by the EH. Our empirical analysis focuses on monthly U.S. Treasury bond excess returns ($\tau = 1$) over the period 1962 to 2018 constructed using the Gürkaynak, Sack, and Wright (2007) dataset and a one-month Treasury bill obtained from the Center for Research in Security Prices (CRSP) as in Gargano et al. (2019).[8] By using monthly holding period returns, we avoid the issues that the artificial persistence induced from using annual overlapping returns can have on inference procedures (Bauer and Hamilton, 2018). Additionally, the higher return frequency allows us to better capture short-lived dynamics in bond excess returns across economic states (Farmer et al., 2019, Gargano et al., 2019).

[Insert Figure 1 About Here]

Figure 1 plots time series of excess returns for Treasury bonds with two, three, four, and five years to maturity, respectively.[9] Bond excess returns are notably more volatile during the early 1980s, smaller and less variable since the 1990s, but visibly increasing in size and variability during the 2000 and 2008 recessions.

---

[8]We detail the construction of monthly log yields and bond prices in the Internet Appendix. The data are available at https://www.federalreserve.gov/data/nominal-yield-curve.htm.

[9]The same set of maturities are considered in, e.g., Fama and Bliss (1987), Cochrane and Piazzesi (2005), Ludvigson and Ng (2009), Thornton and Valente (2012), Eriksen (2017), and Gargano et al. (2019).

[Insert Table 1 About Here]

Panel A of Table 1 presents descriptive statistics for monthly bond excess returns. Mean excess returns and volatilities are increasing with maturity, whereas Sharpe ratios are declining from a high of 0.46 for the two-year bond to 0.35 for the five-year bond. Short-term bonds exhibit higher skewness, kurtosis, and have somewhat more persistent excess returns. Importantly, first-order autocorrelation coefficients for monthly bond excess returns are substantially lower than those typically observed in studies using annual overlapping bond excess returns (see, e.g., Cochrane and Piazzesi (2005) and Ludvigson and Ng (2009)) and the first-order autocorrelation coefficient never exceeds 0.17 across the maturity spectrum. Panel B of Table 1 provides contemporaneous bond excess return correlation across maturities and confirms the well-known observation that bond excess returns are strongly cross-sectionally correlated across maturities.

B.   Predictor variables

Our empirical analysis centers on out-of-sample forecasts generated using predictive models as in (1) with a collection of standard bond return predictors proposed in the literature. Specifically, we consider (i) yields spreads (Campbell and Shiller, 1991, CS) computed as the difference between the yield on a bond with k months to maturity and the implied yield on a one-month Treasury bill obtained from CRSP, (ii) forward spreads (Fama and Bliss, 1987, FB) computed as the difference between the k-month forward rate and the one-month yield, (iii) the first three principal components (PC) of yields (Litterman and Scheinkman, 1991) obtained using 12, 24, 36, 48, and 60 month yields, (iv) a linear combination of forward rates (Cochrane and Piazzesi, 2005, CP) obtained from projecting 12, 24, 36, 48, and 60 month forward rates onto the mean excess bond return across the maturity spectrum, and (v) a linear combination of macroeconomic factors (Ludvigson and Ng, 2009, LN) obtained using

11

the FRED-MD database ([McCracken and Ng, 2016](#)) and estimated analogously to CP.[10] All variables are constructed recursively in the out-of-sample exercise.[11] Section [IA.C.2](#) in the Internet Appendix provides additional details on variable construction and descriptive statistics.

C. State variables

Traditional bond return predictability tests ask whether bond excess returns are predictable *on average*, not *if and when* bond returns are predictability. We address this issue by introducing a new method to gauge periodic, state-dependent predictability in a multivariate setting that rests on the basic premise that even unconditionally poor predictors may sharpen bond risk premia predictions in certain states of the world. To uncover such states, we need to identity state variables that are likely to capture fluctuations in forecast losses. We consider two variables well-known to the literature.[12]

The first state variable we consider is the Purchasing Managers' Index (PMI) published by the Institute of Supply Management. PMI is a closely watched barometer of business conditions released on the first business day of every month and is regarded as a prime leading indicator of the business cycle ([Berge and Jordà, 2011](#), [Christiansen et al., 2014](#)).[13]

---

[10]The literature offers more variables for utilizing macroeconomic information in prediction and these could naturally be entertained as well. We choose to focus on a standard set here, but note that variables such as the output gap ([Cooper and Priestley, 2009](#)) and interest rate cycles ([Cieslak and Povala, 2015](#)) similarly reveal predictable state-dependencies. Their inclusion does not alter our findings, if anything, our results become stronger.

[11][Ghysels et al. (2018)](#) argue that the predictive power of macro variables for future overlapping excess bond returns originates largely from data revisions not available to forecasters in real-time (see also [Fulop, Li, and Wan (2020)](#)). [Eriksen (2017)](#) and [Huang, Jiang, Tong, and Zhou (2020)](#), on the other hand, show that carefully identifying the data available at the time of the forecast using vintages of real-time data recovers most of this predictability.

[12]We provide additional details in Internet Appendix [IA.C.3](#) along with descriptive statistics.

[13]Bloomberg, for example, offers a world map of business conditions based on PMI here:

Using a variable that tracks business cycle fluctuations to assess state-dependencies in bond predictability is motivated by a natural link to time-varying risk aversion (Campbell and Cochrane, 1999, Wachter, 2006, Creal and Wu, 2020) and a large literature that documents differences in predictability connected to the business cycle (Rapach et al., 2010, Henkel et al., 2011, Dangl and Halling, 2012, Andreasen et al., 2020).

The second state variable is the macroeconomic uncertainty index ($\mathcal{U}$) proposed in Jurado et al. (2015).[14] Macroeconomic uncertainty has recently been identified as an important contributor to business cycle fluctuations (Bloom, 2009, Ludvigson, Ma, and Ng, 2019) and asset prices (Drechsler, 2013, Bali, Brown, and Tang, 2017, Borup and Schütte, 2020). Moreover, it has recently been used to study state-dependent performance of affine term structure models (Sarno et al., 2016). Last, uncertainty is likely to be linked to risk aversion (Bekaert, Engstrom, and Xu, 2019), which bears direct influence on the required compensation for bearing interest rate risk.

[Insert Figure 2 About Here]

Figure 2 displays the evolution of the state variables over time. To identify states and to facilitate interpretation, we define periods of high (low) activity and uncertainty, respectively, using the 80% (20%) quantiles of their time series. These states are represented by green (yellow) shaded areas in the graph. We refer to the remaining periods as normal activity and uncertainty states, respectively. The quantiles are chosen to represent extreme states, while ensuring a reasonable number of observations for inference.[15] PMI and $\mathcal{U}$ both identify

---

https://www.bloomberg.com/graphics/global-pmi-tracker/.

[14]We focus on the index associated with $h = 1$ step ahead forecast errors to match the holding period of the bond as well as the data frequency in general. The index is available from Sydney C. Ludvigson's website: https://www.sydneyludvigson.com/macro-and-financial-uncertainty-indexes.

[15]Similar approaches to identifying high/low and good/bad states can be found in, among others, Liew and Vassalou (2000), Rapach et al. (2010), and Sarno et al. (2016). We note that our definition of low activity is broader than simply a recession, which would roughly correspond to using a 10% quantile.

13

persistent states, where PMI ($\mathcal{U}$) mostly takes on low (high) values in bad times. The two series have a sample correlation of -0.48, suggesting that the series are related, but not perfect substitutes.

III.   Testing for state-dependent predictability

In this section, we develop a multivariate test for equal conditional predictive ability that allows us to assess and identify state-dependencies in predictability in real-time. We further propose a simple forecast combination scheme with dynamic trimming of methods designed to capitalize on predictable state-dependencies. The method nests the standard equal-weighted combination among all forecasts as well as dynamic method selection.

A.   Notation

To introduce a general notation, let $\boldsymbol{w}_t \equiv (y_t, \boldsymbol{x}_t)'$ be an observed vector defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where y_t is the target object of interest and $\boldsymbol{x}_t$ is a vector of predictors. We consider a setting in which $p + 1$, $p \geq 1$, methods are available for forecasting τ periods into the future. We denote the forecast of $y_{t+\tau}$ constructed at time t by $\widehat{f}^i_{t+\tau} = f^i \left(\boldsymbol{w}_t, \boldsymbol{w}_{t-1}, \ldots, \boldsymbol{w}_{t-m^i+1}; \widehat{\boldsymbol{\theta}}^i_{t,m^i} \right)$ for $i = 1, \ldots, p+1$, where f^i is a measurable function. $\widehat{\boldsymbol{\theta}}^i_{t,m^i}$ denotes the parameter estimates used to construct the forecast for the ith forecasting method obtained using observations from the m^i most recent periods. For ease of exposition, and along the lines of Giacomini and White (2006), we define $m = \max \{m^1, \ldots, m^{p+1}\}$ and require that $m < \infty$. This allows for rolling window estimators, but excludes expanding window forecasting schemes from our test. The number of out-of-sample forecasts is $T = N - (m + \tau - 1)$ for a total sample size of N (time series) observations. In order to assess the forecasting ability of each forecasting method, we use a real-valued loss function $L_{t+\tau} \left(Y_{t+\tau}, \widehat{f}^i_{t+\tau} \right)$. Important examples of L include economic measures such as utility or profits (Granger and Machina, 2006) and statistical measures such as the square or absolute value of the forecast errors

([West](), [2006]), where forecast errors are given by $e_{t+\tau}^i = \widehat{f}_{t+\tau}^i - y_{t+\tau}$. To ease notation, we suppress the arguments of L and write the ith loss function as $L_{t+\tau}^i$ for the remainder of the paper.

B. Rolling window forecasts

The empirical analysis is based on out-of-sample forecasts generated by predictive regression models as in (1).[16] We consider a set of $p+1$ methods, indexed by i, defined by the p predictors outlined in Section B in addition to the EH benchmark. We estimate the predictive regression models by a rolling window OLS scheme, in accordance with our assumptions, and generate forecasts at time t according to (suppressing maturity-dependence for notational simplicity)

$$(2) \qquad \widehat{f}_{t+\tau}^i = \widehat{\alpha}_t^i + \widehat{\boldsymbol{\beta}}_t^i \boldsymbol{x}_t^i,$$

for $i = 1, \ldots, p$ with $\widehat{\boldsymbol{\theta}}_{t,m^i}^i = \left(\widehat{\alpha}_t^i, \widehat{\boldsymbol{\beta}}_t^i \right)'$. The benchmark EH forecast naturally includes no predictors and is simply defined as $\widehat{f}_{t+\tau}^{p+1} = \widehat{\alpha}_t^{p+1}$, which is consistent with a no-predictability interpretation implied by financial theory.

C. The hypothesis of equal conditional predictive ability

We are interested in formally evaluating whether a set of $p+1$, $p \geq 1$, forecasting methods display equal conditional predictive ability using some information set (σ-field), $\mathcal{G}_t$. That is, we wish to test the null hypothesis that

$$(3) \qquad \mathbb{H}_0 \colon \mathbb{E}\left[L_{t+\tau}^i | \mathcal{G}_t \right] = \mathbb{E}\left[L_{t+\tau}^{i+1} | \mathcal{G}_t \right], \quad i = 1, \ldots, p,$$

---

[16]We emphasize, however, that our econometric framework is not limited to such regressions, but naturally includes a broader class of parametric, non-parametric, and Bayesian methods.

or, equivalently, that

$$\mathbb{H}_0 \colon \mathbb{E}\left[\Delta \boldsymbol{L}_{t+\tau} | \mathcal{G}_t\right] = \boldsymbol{0}, \tag{4}$$

where $\Delta \boldsymbol{L}_{t+\tau} = \left(\Delta L^1_{t+\tau}, \ldots, \Delta L^p_{t+\tau}\right)'$ is a p-dimensional vector of loss differentials (i.e. $\Delta L^i_{t+\tau} = L^i_{t+\tau} - L^{i+1}_{t+\tau}$ for $i = 1, \ldots, p$ with $L^i_{t+\tau}$ being the loss function for the ith method). The null hypothesis in (4) is empirically relevant for a number of financial and macroeconomic applications and offers at least three main advantages.[17] First, it allows us to study not only *if* but also *when* there are detectable differences in the predictive accuracy of two or more methods. This is distinctly different from the traditional approach of assessing equal predictive accuracy *on average* and may facilitate the discovery of predictive ability in certain states of the world as captured by $\mathcal{G}_t$. An equivalent viewpoint is that the null hypothesis implies that $\mathcal{G}_t$ is uninformative about the relative predictive accuracy of one or more forecasting methods when forecasting the object of interest τ periods into the future, whereas a rejection implies that the relative predictive accuracy is predictable by $\mathcal{G}_t$ and potentially exploitable in a real-time forecasting environment. Second, if $\mathcal{G}_t$ is set to the trivial σ-field, $\mathcal{G}_t = \{\emptyset, \Omega\}$, then the null hypothesis reduces to an unconditional test comparable to that considered in Mariano and Preve (2012) that provides information about the average predictive ability of the forecasting methods as in Diebold and Mariano (1995) and West (1996). Third, the loss functions depend explicitly on the parameter estimates and not on their probability limits, leading to a test statistic that takes into account estimation uncertainty. Importantly, by allowing for asymptotically non-vanishing estimation uncertainty, the test can accommodate the empirically relevant case of nested models in the set of forecasting methods.[18] This is

---

[17]The null nests the special case of studying predictability across recessions and expansions (see, e.g., Henkel et al. (2011) and Dangl and Halling (2012) for stocks and Gargano et al. (2019), Andreasen et al. (2020), and Bianchi et al. (2020) for bonds), but is more general and allows for continuous state variables.

[18]Technically, with $\mathcal{G}_t = \{\emptyset, \Omega\}$ and asymptotically vanishing estimation uncertainty, the standard errors of differences in forecast performance between a set of nested models will equal zero, leading to non-standard

key in our context as the EH model is nested within all methods coming from the predictive framework in (1).

D.   The multivariate test statistic

The null hypothesis in (4) can equivalently be stated as

$$(5) \qquad \mathbb{H}_0 \colon \mathbb{E}\left[\widetilde{h}_t \Delta \boldsymbol{L}_{t+\tau}\right] = \boldsymbol{0}$$

for all $\mathcal{G}_t$-measurable functions $\widetilde{h}_t$ and $\mathcal{F}_t \subseteq \mathcal{G}_t$. We restrict attention to a subset of these functions collected in the q-dimensional vector $\boldsymbol{h}_t = \left(\widetilde{h}_t^{(1)}, \ldots, \widetilde{h}_t^{(q)}\right)'$. We refer to this vector as a state function. Given a state function, we can then reformulate the multivariate null hypothesis of equal conditional predictive ability as follows

$$(6) \qquad \mathbb{H}_{0,h} \colon \mathbb{E}\left[\boldsymbol{h}_t \otimes \Delta \boldsymbol{L}_{t+\tau}\right] = \boldsymbol{0},$$

where the subscript h indicates the dependence on the state function and $\otimes$ denotes the Kronecker product. The specification in (6) is a natural multivariate extension of the null hypothesis considered in Giacomini and White (2006) and we indeed obtain their econometric framework as a special case for $p = 1$.

As is common in the bond return predictability literature, we consider one-step ahead predictions, $\tau = 1$, as our leading example throughout, but provide theoretical results for multi-step ahead forecasting, i.e. $\tau > 1$, in the Internet Appendix along with our assumptions that are adopted from Giacomini and White (2006). The information set $\mathcal{G}_t$ contains the state variables discussed in Section C. Finally, let $\boldsymbol{d}_{t+1} = \boldsymbol{h}_t \otimes \Delta \boldsymbol{L}_{t+1}$ and consider the quadratic

---

limiting distributions of the test statistics.

statistic

$$(7) \qquad S_h = T \overline{\boldsymbol{d}}' \widehat{\boldsymbol{\Sigma}}_T^{-1} \overline{\boldsymbol{d}},$$

where $\overline{\boldsymbol{d}} \equiv T^{-1} \sum_{t=1}^{T} \boldsymbol{d}_{t+1}$ and $\widehat{\boldsymbol{\Sigma}}_T \equiv T^{-1} \sum_{t=1}^{T} \boldsymbol{d}_{t+1} \boldsymbol{d}'_{t+1}$ is a $(qp \times qp)$ sample covariance matrix that consistently estimates the variance of $\boldsymbol{d}_{t+1}$.[19] That is, S_h is a natural Wald statistic constructed for testing whether $\overline{\boldsymbol{d}}$ is a zero vector. When formulating an alternative hypothesis, it is important to take into account that the data may exhibit non-stationarity. We provide a discussion in the Internet Appendix. For some $c > 0$, we formulate the alternative in line with Giacomini and White (2006) as

$$(8) \qquad \mathbb{H}_{A,h} \colon \mathbb{E}\left[\overline{\boldsymbol{d}}'\right] \mathbb{E}\left[\overline{\boldsymbol{d}}\right] \geq c,$$

for all T sufficiently large. Under stationarity, the null and alternative hypothesis are exhaustive. Under non-stationarity, this is not necessarily the case. If an important $\mathcal{G}_t$-measurable variable is omitted from the state function, it may happen that $\mathbb{E}\left[\overline{\boldsymbol{d}}'\right] \mathbb{E}\left[\overline{\boldsymbol{d}}\right] = 0$ for a particular sample size due to, for instance, shifting means without the null hypothesis being true. As an example, one could easily imagine a situation where one method outperforms (some of) the other methods in certain states, while it performs worse than those methods in other states. Therefore, the test has little power against alternatives where the loss differentials are correlated with $\mathcal{G}_t$-measurable random variables not included in the state function. While this concern is important, it also highlights the flexibility of the test statistic. As mentioned above, the econometrician chooses the state function to include state variables relevant for disentangling the forecasting abilities of two or more forecasting methods. The

---

[19] We note that for large values of q and/or p, the dimension of $\boldsymbol{\Sigma}_T$ and $\overline{\boldsymbol{d}}$ may become large, potentially leading to issues with statistical inference in finite samples. We propose remedies in Borup and Thyrsgaard (2017), but note that our empirical analyses imply reasonable dimensions with one state variable and $p = 5$ predictors plus the EH benchmark.

test, therefore, only provides power in situations when this is possible. As a result, the test statistic changes with the choice of state function, which is emphasized through the subscript in S_h.

The asymptotic properties of the test statistic are summarized in Theorem 1 and we provide proofs and derivations in the Internet Appendix.

Theorem 1 (**One-step multivariate conditional predictive ability test**). *Suppose that Assumptions 1-3 hold (see the Internet Appendix for details). Then the test statistic has the following properties.*

A. ***Asymptotic distribution under the null***. *For forecast horizon $\tau = 1$, state function sequence $\{\boldsymbol{h}_t\}$, $m < \infty$, and under $\mathbb{H}_0$ in (4),*

$$(9) \qquad\qquad S_h \xrightarrow{d} \chi^2(qp), \quad as \ T \to \infty.$$

B. ***Consistency under the alternative***. *For any $c \in \mathbb{R}_+$ and under $\mathbb{H}_{A,h}$ in (8),*

$$(10) \qquad\qquad \mathbb{P}[S_h > c] \to 1, \quad as \ T \to \infty.$$

C. ***Permutation invariance***. *Let $\boldsymbol{L}_{t+1}^*$ be an arbitrary permutation of the vector of forecast losses, and define $\Delta \boldsymbol{L}_{t+1}^* = \boldsymbol{D}\boldsymbol{L}_{t+1}^*$, where*

$$(11) \qquad\qquad \boldsymbol{D} = \begin{bmatrix} 1 & -1 & 0 & \dots & 0 \\ 0 & 1 & -1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & 1 & -1 \end{bmatrix}$$

is a $p \times (p+1)$ matrix. Let $\overline{\boldsymbol{d}}^* = T^{-1}\sum_{t=1}^{T} \boldsymbol{d}_{t+1}^*$ with $\boldsymbol{d}_{t+1}^* = \boldsymbol{h}_t \otimes \Delta \boldsymbol{L}_{t+1}^*$ and $\widehat{\boldsymbol{\Sigma}}_T^* \equiv$

$\frac{1}{T} \sum_{t=1}^{T} \boldsymbol{d}_{t+1}^* \boldsymbol{d}_{t+1}^{*'}$. *Then,*

$$(12) \qquad\qquad S_h^* \equiv T \overline{\boldsymbol{d}}_m^{*'} \left(\widehat{\boldsymbol{\Sigma}}_T^*\right)^{-1} \overline{\boldsymbol{d}}_m^* = S_h, \quad \forall T.$$

We provide a corresponding result for the unconditional, possibly multi-step, case, in the Internet Appendix. The unconditional case, in which we compare the average performance of the methods over the out-of-sample window, is obtained by setting $\boldsymbol{h}_t = 1$ for all t. The limiting distribution is then $\chi^2(p)$ for a test statistic that employs a HAC type covariance matrix estimator to capture serial dependence. In the case of the conditional test and multi-step forecast horizons, a $\chi^2(qp)$ limiting distribution is obtained when using an appropriate HAC type covariance matrix estimator. Although any reordering of the forecasting methods alters the dynamics of $\boldsymbol{d}_{t+1}$, Theorem 1.C. shows that we obtain the same value of the test statistics and the same limiting distribution under the null hypothesis for each permutation (reordering) of the forecasting methods, regardless of the null being true or not. This is important as it allows the researcher to perform just a single test. Appendix IA.B provides Monte Carlo simulations confirming proper small sample properties of the test in a setting that matches the empirical application below.

E. Understanding the test

This section provides an intuitive illustration of the test statistic. Consider the simple case of $p = 1$, where the problem reduces to a comparison between a single pair of forecasting methods. An unconditional test is equivalent to the regression

$$(13) \qquad\qquad \Delta L_{t+1} = \varphi_0 + \eta_{t+1},$$

where the null hypothesis that $\varphi_0 = 0$ can be tested using a standard t-test using an appropriate HAC type of covariance estimator to account for serial correlation under the

null hypothesis. The conditional test augments the regression with a set of state variables. Suppose that we consider a single state variable $\tilde{h}_t$ (as in the empirical analysis below to facilitate economic interpretation), then the conditional test amounts to running the extended regression

$$(14) \qquad \Delta L_{t+1} = \boldsymbol{\varphi} \boldsymbol{h}_t + \eta_{t+1} = \varphi_0 + \varphi_1 \tilde{h}_t + \eta_{t+1},$$

with $\boldsymbol{\varphi} = (\varphi_0, \varphi_1)$ and $\boldsymbol{h}_t = \left(1, \tilde{h}_t\right)'$ being the state function.[20] In this case, we are interested in testing jointly $\varphi_0 = \varphi_1 = 0$ using a Wald test and an appropriate estimator of the covariance matrix. The limiting distribution under the null hypothesis is equivalent to the ones provided in Theorem 1. From (14), it is clear that a rejection of $\varphi_1 = 0$ indicates that the state variable $\tilde{h}_t$ is informative about the future relative predictive ability of the methods under consideration. That is, there is evidence of state-dependency. Importantly, the expression in (14) is nothing more than a full sample predictive regression similar in spirit to (1) estimated over the out-of-sample window. The key difference is that (14) predicts the future relative predictive ability among the candidate forecasting methods using state variables whose values are observable at the construction of the forecast and are picked by the researcher. We refer to them as state regressions in the following. These ideas naturally extend to our case of $p > 1$, in which case the test statistic is similar tox a seemingly unrelated regression (SUR) framework. We will make use of this insight below when formulating a simple decision rule to exploit rejections of the null hypothesis to dynamically select or combine among forecasting methods with indistinguishable predicted performance.

---

[20]Using multiple state variables, in addition to the constant, is then equivalent to running a multiple regression and conducting joint inference on all parameters.

F. Ranking of forecasting methods

A rejection of the null hypothesis suggests that one or more forecasting methods exhibit superior predictive ability in certain states, but provides little guidance towards identifying the method(s) causing the rejection and the predictive state. This identification is of both economic and practical interest. Central banks, international organizations (IMF, OECD, and the World Bank), and professional forecasters (SPF and Blue Chip) frequently generate forecasts that are widely followed by market participants and policy makers. Designing routines that can identify when forecasts and/or forecasters are predicted to do well in real-time is clearly relevant.

We therefore propose a dynamic rule that facilitates the identification of the best set of methods in real time conditional on the realization of one or more state variables at the time of the forecast. The procedure is designed to reveal potential fluctuations in predictive ability over time, similar in spirit to the fluctuation test of Giacomini and Rossi (2010), and their origins through the use of observable state variables. In formulating the rule, we consider a MCS-type procedure (Hansen et al., 2011) that eliminates inferior methods according to a prespecified elimination rule and rank forecasting methods into a best set whose elements have equal predictive ability conditional on the information available at the time of the forecast.

To identify the best set, we divide the out-of-sample window into two parts. The first part has length T_1 and is used for initially estimating the state regression. The second part has length T_2, with $T_1 + T_2 = T$, over which the dynamic procedure will be employed. We introduce the following additional notation for the dynamic ranking rule. First, for every time t we let M_t denote the set of forecasting methods available at time t. The goal of the proposed procedure is to eliminate any inferior forecasting methods from this set, leaving the set M_t^* of superior forecasting methods whose predictive ability are indistinguishable. The rule is formulated using a single state variable (and a constant), but it can be extended

directly to a setting with multiple state variables. We then propose the following three-step ranking algorithm at each time point $t = m + T_1, \ldots, N - 1$.

Step 0: For all pairwise combinations of forecasting methods, $j, i \in M_t, i \neq j$ estimate by OLS the regression model

$$(15) \qquad \Delta L_{s+1}^{(i,j)} = L_{s+1}^i - L_{s+1}^j = \boldsymbol{\varphi}^j \boldsymbol{h}_s + \eta_{s+1}$$

with $s = t - T_1 + 1, \ldots, t$, i.e. using a rolling window of the past T_1 forecast errors. The conditional expectation $\mathbb{E}\left[\Delta L_{t+1}^j | \mathcal{G}_t\right]$ is estimated by $\widehat{\boldsymbol{\varphi}}^j \boldsymbol{h}_t = \widehat{\varphi}_0^j + \widehat{\varphi}_1^j \widetilde{h}_t$, which measures the time t prediction of the future relative performance of method i and j given the current information in the state variable. Based on those predictions, rank all $p + 1$ methods (using a normalization of one method). The forecasting method with lowest predicted loss across all pairwise combinations is ranked first and the method with highest predicted loss is ranked last.

Step 1: Run the multivariate test for equal conditional predictive ability.

Step 2: If the test is not rejected, set $M_t^* = M_t$. Otherwise, eliminate the lowest ranked forecasting method from M_t based on the ranking of predicted forecast losses. Iterate Steps 1–2 until the null is no longer rejected.

A few remarks are in order. First, the ranking rule exploits the state regression interpretation of our test statistic and is strongly rooted in econometric theory. Second, since the elimination of models is based on a state-specific ranking, the ranking rule will capture the state-dependency of predictability over time. Third, the set of superior forecasting methods at any given time t will not only depend on the states used, but also on the choice of loss function. This leaves the researcher with some flexibility in choosing how different types of forecast errors should be weighted. Fourth, since the algorithm provides sets of equal predictive ability within each state, the ranking rule can be viewed as a conditional MCS

algorithm. Fifth, since the test is permutation invariant (see Theorem 1), the test only needs to be carried out once each time Step 2 is conducted, even when elimination alters the ordering of the methods. Sixth, the proposed method is forward looking in the sense that M_t^* is comprised of the methods with lowest expected future loss. This sets our method apart from others previously considered in the literature (see, e.g. Aiolfi and Timmermann (2006) and Adämmer and Schüssler (2020)), which have focused solely on ranking methods based on their past performance. Ranking solely on past, or even the most recent, performance is especially problematic if the performance of the forecasting methods display abrupt shifts, e.g. according to states, which our empirical evidence suggests is the case when it comes to forecasting the bond risk premium. Finally, considering the constituents of M_t^* across time or states informs about state-dependency in predictability, which we make use of in the empirical section below to understand the success of this methodology and derive economic insight.

G. Forecast combination with dynamic trimming

The set of best methods, M_t^*, may consist of a singleton, some methods, or all methods, which suggests that the elimination rule can be interpreted as a way to dynamically trim the set of candidate predictors. Accordingly, we propose a simple dynamically combined forecast $\widehat{f}_{t+1}^*$ defined as

$$(16) \qquad \widehat{f}_{t+1}^* = \frac{1}{|M_t^*|} \sum_{i \in M_t^*} \widehat{f}_{t+1}^i,$$

where $|M_t^*|$ denotes the number of elements in M_t^*. This definition is consistent with the simple equal-weighting principle often found in the literature, but with the modification that we dynamically trim the set of methods prior to combination based on the predicted

losses from the dynamic ranking rule.[21] In that way, this methodology allows for selecting a single method, combining among all methods, or any number of methods in between dynamically in real time. Our approach can therefore be interpreted as a data-driven and forward-looking way to select between dynamic model selection and forecast combination. We refer to this strategy as *dynamic forecast combination.* The strategy complements the static combination and trimming rules found, for instance, in Rapach et al. (2010) and provides a formal routine to identify the optimal set of methods prior to combination (Aiolfi et al., 2011).[22] It also extends the procedure suggested in Adämmer and Schüssler (2020) by combining among methods that are predicted to have the smallest future losses via the use of (15) in Step 0, as opposed to those that have the smallest past losses. Moreover, our selection of methods is rooted in a formal multivariate testing procedure, resembling a conditional MCS. In the special case with only two methods, $p = 1$, our scheme reduces to the switching rule provided in Giacomini and White (2006). Timmermann and Zhu (2017) formally show that forecast improvements are guaranteed when state variables are powerful and Granziera and Sekhposyan (2019) provide empirical evidence consistent with this observation.

IV.   State-dependencies in bond return predictability

This section presents empirical evidence on predictable state-dependencies in bond excess return predictability. While individual bond predictors fail to consistently deliver reliable

---

[21]The equal-weighted combination scheme has a long tradition in the forecasting literature and is empirically hard to beat as it involves no estimation error in weights (Timmermann, 2006, Rapach et al., 2010). Other combination schemes are naturally possible, e.g. using estimated least squares weights, possibly with shrinkage to equal weights (Bates and Granger, 1969, Granger and Ramanathan, 1984, Zellner, 1986, Diebold and Pauly, 1987). While one could possibly increase forecast performance further by considering more complicated combination schemes, this is not the aim of our paper.

[22]Alternative suggestions include determining the optimal set based on past performance (Aiolfi and Timmermann, 2006), the (unconditional) model confidence set (Samuels and Sekkel, 2017), and lasso-based procedures (Diebold and Shin, 2019).

out-of-sample forecasts on average, we find strong evidence for state-dependent predictability related to economic activity and uncertainty using the full sample test for equal conditional predictability ability. Exploiting these state-dependencies using a dynamic forecast combination strategy leads to substantial gains in forecast accuracy. The empirical findings are consistent with the EH holding periodically.

A. Out-of-sample predictability

We begin by gauging the unconditional predictive ability of individual bond predictors using a rolling window estimation scheme. We use the period January 1962 to December 1989 as our initial in-sample estimation period, similarly to Gargano et al. (2019), leaving the remaining period for out-of-sample predictions. The period January 1990 to December 1999 is set to our initial testing period, used in the dynamic ranking rule, and the period from January 2000 to December 2018 as our evaluation period. We focus on U.S. Treasury bonds with 24, 36, 48, and 60 months to maturity and consider models based on the predictor variables outlined in Section B. To evaluate the out-of-sample performance of the predictive methods relative to the constant expected return benchmark implied by the EH, we consider the Campbell and Thompson (2008) out-of-sample R^2 statistic

$$
(17) \qquad R^2_{OS,i,k} = 1 - \frac{\sum_{t=R+1}^{N} \left(rx_t^{(k)} - \widehat{rx}_{t,i}^{(k)} \right)^2}{\sum_{t=R+1}^{N} \left(rx_t^{(k)} - \widehat{rx}_{t,EH}^{(k)} \right)^2},
$$

where $\widehat{rx}_{t+1,i}^{(k)}$ and $\widehat{rx}_{t+1,EH}^{(k)}$ denote the forecast from the ith predictor model and the EH benchmark, respectively, $R = m + T_1$ denotes the end of the testing period, and N denotes the total number of observations.[23] The R^2_{OS} statistic in (17) is equivalent to one minus the ratio of mean squared prediction errors, i.e. $R^2_{OS,i,k} = 1 - \frac{MSPE_i^{(k)}}{MSPE_{EH}^{(k)}}$. Accordingly, we employ the squared prediction error in place of L_{t+1} when implementing our statistical tests on predictive ability. An $R^2_{OS} > 0$ implies that the MSPE of the ith predictor model is lower than that of

---

[23]See also Fama and French (1989) for a similar definition and use.

the EH benchmark model, indicating higher predictive accuracy. We interpret the EH model as a no-predictability benchmark and test the null of no predictability $\left(R^2_{OS} \leq 0\right)$ against the one-sided alternative of predictability by the ith predictor model $\left(R^2_{OS} > 0\right)$ using the Diebold and Mariano (1995) (DM) test for equal predictive ability.[24]

[Insert Table 2 About Here]

Panel A of Table 2 reports R^2_{OS} values and DM p-values for individual bond predictors across the maturity spectrum. The main conclusion is that no individual model is able to reliably outperform the EH benchmark unconditionally for all maturities. The majority of models deliver negative R^2_{OS} values and those that are positive are far from being significant at conventional levels.[25] These results are in line with Gargano et al. (2019), who similarly find few positive R^2_{OS} values for linear predictive models.[26] Panel B considers a simple (static) equal-weighted forecast combination scheme (Bates and Granger, 1969, Timmermann, 2006, Rapach et al., 2010). We denote this combined forecast by EW. The combined forecast generates positive R^2_{OS} values from 6.08% for the two-year bond to 4.58% for the five-year bond. These values are all significant according to the DM p-value at the five percent level. That is, although no individual predictor is able to consistently outperform the EH, a simple equal-weighted average of the individual forecasts is.

[Insert Figure 3 About Here]

Figure 3 plots the cumulative difference in squared prediction errors (CDSPE) between

---

[24]Note that this is the unconditional version of the test statistic in Giacomini and White (2006) which is nested within our framework for $p = 1$.

[25]We provide in-sample predictive regression results in the Internet Appendix, where we show that our set of predictors are reliably related to bond risk premia.

[26]Our rolling window regression results for CP is poorer than in Gargano et al. (2019), indicating that bond return predictability is sensitive to the forecasting setup. In unreported results, we indeed find that many R^2_{OS} values improve when considering a forecasting environment with an expanding window instead. However, the main conclusions remain qualitative similar.

the EH and the ith predictor model

$$(18) \qquad \text{CDSPE}_{t,i}^{(k)} = \sum_{l=R+1}^{t} \left(rx_l^{(k)} - \widehat{rx}_{l,EH}^{(k)} \right)^2 - \sum_{l=R+1}^{t} \left(rx_l^{(k)} - \widehat{rx}_{l,i}^{(k)} \right)^2,$$

where $R+1$ denotes the time of the first forecast, and $\widehat{rx}_{t+1,i}^{(k)}$ and $\widehat{rx}_{t+1,EH}^{(k)}$ denote the forecast from the ith predictor model and the EH benchmark, respectively. This graphical device is suggested by Goyal and Welch (2008) as a way to assess relative performance over time (and is indirectly a visual inspection of state-dependencies). Figure 3 plots the CDSPEs against economic activity and uncertainty states identified using PMI (left colum) and $\mathcal{U}$ (right commmn), respectively, to assess the relation between relative forecasting performance and the state variables. The plots supports the use of conditioning variables that tracks salient features of the business cycle as several swings in CDSPE for the individual predictors is related to states associated with high and low values of economic activity and uncertainty, respectively.

B. Testing for equal conditional predictive ability

The previous section establishes that linear predictive methods are unable to reliably beat the EH on average. However, this does not exclude the possibility that some methods provide significantly better forecast in certain states of the world. We investigate this possibility using the multivariate test for equal conditional predictive ability introduced in Section III. We entertain three specifications for the state regression. First, we consider the information in PMI to examine if predictive ability is related to economic activity and specify the state function as $\boldsymbol{h}_t = (1, \text{PMI}_t)'$. Second, we specify $\boldsymbol{h}_t = (1, \mathcal{U}_t)'$ to study the effect of macroeconomic uncertainty.[27] Third, we consider an unconditional version of the multivariate

---

[27] In principle, one could also combine the state variables into $\boldsymbol{h}_t = (1, \text{PMI}_t, \mathcal{U}_t)'$ to exploit their joint information. We leave the analysis of multivariate state regressions for further research and instead focus on univariate state regressions here as they more easily facilitates state-wise interpretation.

test in which we set $\boldsymbol{h}_t = 1$ for all t and denote the case by NONE.

[Insert Table 3 About Here]

Table 3 reports test statistics and corresponding p-values for the multivariate test for equal (un)conditional predictive ability over the evaluation period for each of the three specifications for the state regression. The implementation is based on a sample covariance matrix as dictated by theory (see Section III and the Internet Appendix).[28] The strong rejections of the null hypothesis of equal conditional predictive ability across all maturities for both specifications of $\boldsymbol{h}_t$ that uses conditioning information provides substantial evidence favoring state-dependencies in bond excess return predictability. The unconditional test, on the other hand, fails to reject equal predictive abilities across all maturities. In other words, the state variables enable the detection of conditional differences in predictive accuracy. We conclude that individual predictors are equally accurate when evaluated unconditionally, but statistically different conditional on states of economic activity or uncertainty.

C.   Ranking and elimination of methods

Having established that relative bond return predictability shifts over time with state variables tracking economic activity and uncertainty, we turn to a more detailed analysis of these state-dependencies. We start with an examination of the ranking and elimination of models over the out-of-sample period using the dynamic forecast combination procedure (see Section F). The scheme enables us to identify the best set of methods M_t^* with indistinguishable conditional predictive ability at each point in time.[29]

---

[28]We note that NONE should, in theory, by evaluated using a HAC estimator, but we use a sample estimator here to ease comparison. However, results are both qualitative and quantitatively similar when employing a Newey and West (1987) estimator with a bandwidth of 12 lags.

[29]In Appendix IA.D.5 we depict the size of M_t^* over time using PMI and $\mathcal{U}$. This best set of models varies considerably over time and includes situations in which the set include all models, leading to forecasts equal to EW, and situations with a singleton (equivalent to dynamic method selection). That is, at times there is

Panel C of Table 2 presents the results for the dynamic forecast combination scheme using PMI and $\mathcal{U}$, respectively, as conditioning variables and an unconditional alternative that uses average past predictability as an indicator for future predictability. We label this alternative NONE and consider it a natural and challenging benchmark. This unconditional alternative is related to Samuels and Sekkel (2017) who suggest trimming a given set of models using a recursive implementation of the unconditional MCS and Adämmer and Schüssler (2020) who uses the set of models that, when combined, provides the best past performance. Our conditional alternative trims the set of forecasting methods using a conditional MCS idea with the elimination based on the projected future predictability of bond excess return using current information in state variables. Strikingly, the dynamic forecast combination strategy delivers positive R^2_{OS} values relative to the EH across all conditioning variables and bond maturities that are economically large with values between 5.11% and 7.98% for PMI and between 4.98% and 9.86% for $\mathcal{U}$. Moreover, these values always exceed those of the EW strategy with a considerable margin. All (most) are significant relative to the EH (EW) at conventional levels when using either PMI or $\mathcal{U}$, whereas NONE does not deliver significant improvements against the EW. This suggests that past relative predictive accuracy is not informative about future predictability relative to just using all methods with equal weight. On the other hand, the state variables are remarkably informative about future predictability in the simple manner suggested by our multivariate test and the embedded state regressions, cf. (15).

[Insert Figure 4 About Here]

Figure 4 plots the CDSPE for the dynamic forecast combination strategies and the unconditional alternative NONE relative to the EH. Overall, we find that relative forecasting gains are mostly uniformly distributed across the out-of-sample evaluation period and that

---

no need for trimming of the full set of models and at other times we should only use the forecasts from a single model. Importantly, this tells us why dynamic method selection may lead to improvements over a static forecast combination.

no particular event or period drive the positive results, although we do observe a particularly strong increase during the latest recession relative to the EH benchmark for the five-year bond using PMI as the state variable.

[Insert Figure 5 About Here]

Figure 5 plots the CDSPE for the two dynamic forecast combination strategies and NONE relative to EW. As above, the dynamic forecast combination strategy always performs on par or better than EW. This is also reflected in Panel D of Table 2, where we observe positive R^2_{OS} values that are of economically meaningful magnitudes and most are significant at conventional significance levels. These relative forecasting gains concentrate in periods with low (high) economic activity (uncertainty). That is, our dynamic forecast combination scheme delivers improvements in forecast accuracy in periods of turmoil, exactly when investors arguably needs it the most. Moreover, we see that trimming the set of candidate methods prior to combination using a dynamic rule rooted in a multivariate test for equal conditional predictive ability delivers sizeable improvements. Lastly, the CDSPE curve is almost completely flat for NONE, confirming that static equal-weighting combination performs about as well.

D.   Understanding the sources of conditional predictability

This section studies the underlying sources of conditional predictability and the sizable improvements in predictive accuracy established above. To facilitate interpretation, we classify the continuous state variables, PMI and $\mathcal{U}$, into low, normal, and high economic activity (uncertainty) states using the 20% and 80% quantiles similarly to, e.g., Rapach et al. (2010). We then proceed in two steps. We begin by computing inclusion frequencies for each forecasting method within the low, normal, and high economic activity and uncertainty states, respectively. We then study how the individual methods perform in each state and relate it to the overall performance.

1.   Inclusion frequencies

We compute inclusion frequencies for each forecasting method and state variable using the low, normal, and high states for economic activity (PMI) and uncertainty ($\mathcal{U}$), respectively. Within each state, we then define the inclusion frequency of the ith forecasting method as the fraction of months the model is included in the best set relative to the total number of months in the given state.

[Insert Table 4 About Here]

Table 4 reports a striking pattern in which the EH tends to be eliminated in high (low) economic activity (uncertainty) states across the entire maturity spectrum. Interpreting the EH as a no-predictability benchmark implies that bond excess returns are predictable in high (low) economic activity (uncertainty) states, but less so in other states. The EH, conversely, provides a reliable anchor in periods with low (high) and to some degree normal economic activity (uncertainty). In Appendix IA.D.3, we device a methodology that identifies predictability within states seen over the full out-of-sample period, conditioning on low, normal, and high PMI and $\mathcal{U}$ states, respectively. This strongly supports the striking pattern that EH tends to be eliminated in high (low) economic activity (uncertainty) states across the entire maturity spectrum.

2.   State-dependent predictability

The inclusion frequencies are indicative of when certain models are predicted to do well. In this section, we ask whether the inclusion frequencies align with relative performance. That is, we ask whether the procedure correctly identifies methods with good and bad relative performance.

[Insert Table 5 About Here]

Table 5 reports state-specific R^2_{OS} values for the individual predictors relative to the EH. The results are supportive of the procedure correctly identifying methods that do well.

We find that individual predictors are generally performing poorly ($R^2_{OS} < 0$) in low (high) economic activity (uncertainty) states and well ($R^2_{OS} > 0$) in high (low) economic activity (uncertainty) states.[30] This is consistent with the inclusion frequencies of the EH. Specifically, the procedure appears to correctly anticipate periods in which the EH provides a reasonable anchor for expected bond excess returns and periods in which bond risk premia are predictable. Moreover, there is also a close mapping between the inclusion frequencies and the magnitudes of the R^2_{OS} values, where models are more likely to be included (excluded) in a given state the higher (lower) its R^2_{OS}. That is, the gains in predictive accuracy are coming from the rule's ability to correctly predict predictability.

Corroborating this statement, the dynamic forecast combination schemes using PMI or $\mathcal{U}$ deliver positive and sizable R^2_{OS} values in all states and for all bonds. Moreover, the improvements in accuracy are always comparable or greater than those observed for EW and NONE. This is a central result that highlights the resilience of the dynamic forecasting combination scheme. Even within states where all predictors are weak on average, dynamically combined forecasts performs remarkably well because the scheme correctly trims away methods predicted to perform poorly and retains those predicted to perform well conditional on the realized state variable each period. We obtain similarly impressive performance in states where most predictors are outperforming the EH, which is again attributable to the methods ability to correctly identify which methods to trim and which to retain based on predicted performance.

Overall, our empirical results are consistent with, and clearly points to, state-dependencies in bond excess return predictability linked to economic activity and uncertainty. Bond excess returns are predictable in states with high (low) economic activity (uncertainty), whereas the

---

[30]One exception is for CP in low uncertainty states using $\mathcal{U}$ as state variable. This is mainly due to few, yet notable, observations. Importantly, our procedures are able to capture this fact; within this state, they eliminate CP as the first predictor across all bond maturities in the full out-of-sample ranking and CP possesses lowest (together with EH) inclusion frequency in the dynamic ranking.

EH serves as a reliable anchor in the remaining states of the world. Moreover, these shifts in predictability can be successfully exploited via the methodology proposed in this paper. It is relevant to relate those findings to Gargano et al. (2019) who identify stronger predictability in recessions than outside recessions. Our findings contrast this notion. The main difference is that they use a different set of models and Bayesian methods, incorporating time-varying parameters and/or stochastic volatility, and not the conventional univariate linear models considered in the present paper. Note, however, they do recover the pattern documented in this paper when using FB as the single predictor. Interestingly, our out-of-sample forecasts for bond risk premia remain fully consistent with the results in Gargano et al. (2019) regarding the link to the real economy. We turn to this point below.

V.   Links to the real economy

In this section, we examine the link between our out-of-sample bond risk premia forecasts and the real economy. Standard finance theory implies that investors demand compensation for risks associated with recessions due to heightened risk aversion (Campbell and Cochrane, 1999, Wachter, 2006, Buraschi and Jiltsov, 2007) and risk premia should therefore be countercyclical and peak in recessions.

[Insert Table 6 About Here]

We employ PMI as our measure of economic activity (Berge and Jordà, 2011, Christiansen et al., 2014) and report in Table 6 the contemporaneous correlation among PMI and the risk premia estimates from individual bond predictors, EW, and the dynamically combined forecast generated by PMI, $\mathcal{U}$, and NONE. The results offer two main insights. First, yield-based variables deliver risk premia estimates that are significantly positively correlated with real economic activity. The methods therefore imply procyclical risk premia, which sharply contrasts canonical theory. LN, on the other hand, realizes a significant negative correlation of about -38% across the maturity spectrum, which is consistent with countercyclical risk

premia. These observations are consistent with those of Ludvigson and Ng (2009). The EW combination strategy produces risk premia estimates that are virtually acyclical and unrelated to the state of the economy. The acyclicality is likely caused by the divergent cyclicalities being washed out in the equal-weighted combination scheme. The dynamic combination strategy, which selects individual methods for subsequent combination based on information in the state variables, produces markedly negative and statistically significant correlations with the real economy. Dynamically trimming the candidate set of methods prior to combination based on predicted performance yields *both* economically meaningful risk premia estimates that are strongly countercyclical as well as substantial improvements in overall forecast accuracy.[31]

[Insert Figure 6 About Here]

Supporting this, Figure 6 depicts our dynamic combination forecasts using PMI and $\mathcal{U}$ as state variables, along with recession periods as defined by the NBER. Risk premia display a clear tendency to increase during recessionary periods and decline during expansionary periods, resembling a countercyclicality with business cycles. These findings altogether demonstrate the importance of appropriately selecting among plausible methods as proposed here. Our empirical findings are consistent with recent evidence similarly documenting countercyclical bond risk premia (Gargano et al., 2019, Andreasen et al., 2020, Bianchi et al., 2020) and models with time-varying risk (Bekaert et al., 2009, Carpenter et al., 2020, Creal and Wu, 2020).

---

[31]Other types of business cycle indicators can naturally be entertained. We report in the Internet Appendix contemporaneous correlations among generated forecasts and each of the macroeconomic uncertainty ($\mathcal{U}$), recession probabilities of Chauvet and Piger (2008), the Chicago Fed National Activity Index (CFNAI), and logarithmic growth rates to industrial production growth. It stands out that our dynamic forecasting combination technique leads to much stronger countercyclical bond risk premia than all yield-based variables and EW.

VI. Economic value

This section quantifies the economic value attainable from dynamically combining forecasts across methods predicted to perform well. Specifically, we consider the asset allocation decision of a mean-variance investor with relative risk aversion γ that chooses the weight $\omega_t^{(k)}$ to invest in a k-period bond and the weight $\left(1 - \omega_t^{(k)}\right)$ to invest in a one-period safe bond (Marquering and Verbeek, 2004).[32] The resulting portfolio return is then

$$
(19) \qquad\qquad r_{p,t+1}^{(k)} = y_t^{(1)} + \omega_t^{(k)} r x_{t+1}^{(k)},
$$

where $r x_{t+1}^{(k)}$ denotes monthly bond excess returns for a Treasury bond with k months to maturity. Denoting moments conditional on $\mathcal{G}_t$ by subscript t, we assume that the investor has a utility function $U\left(r_{p,t+1}^{(k)}\right)$ of the form

$$
(20) \qquad\qquad U\left(r_{p,t+1}^{(k)}\right) = \mathbb{E}_t\left[r_{p,t+1}^{(k)}\right] - \frac{1}{2}\gamma \mathrm{Var}_t\left[r_{p,t+1}^{(k)}\right],
$$

where γ is a measure of relative risk aversion. Solving the maximization problem yields the optimal portfolio weights

$$
(21) \qquad\qquad \omega_t^{(k)} = \frac{1}{\gamma}\frac{\mathbb{E}_t\left[r x_{t+1}^{(k)}\right]}{\mathrm{Var}_t\left[r x_{t+1}^{(k)}\right]},
$$

---

[32]Assuming that investors have mean-variance preferences in asset allocation exercises has a long tradition in predictability studies and similar approaches can be found in, among many, Campbell and Thompson (2008), Goyal and Welch (2008), Wachter and Warusawitharana (2009), Thornton and Valente (2012), Sarno et al. (2016), Eriksen (2017), Ghysels et al. (2018), and Gargano et al. (2019). A multivariate asset allocation exercise is, in principle, also possible, yet due to the very high correlation among the considered bonds, cf. Table 1, the inverse of their covariance matrix is highly unstable, leading to unrealistic and very large weights that renders their winzoriation binding in all periods. For these reasons, we do not consider this approach.

where $\mathbb{E}_t\left[rx_{t+1}^{(k)}\right]$ is estimated using the ith predictive method and $\mathrm{Var}_t\left[rx_{t+1}^{(k)}\right]$ is computed using a rolling window of past bond excess return realizations.[33] We winzorize weights according to reasonable shorting and leverage constraints, similarly to Thornton and Valente (2012) and Gargano et al. (2019), such that $\omega_t^{(k)} \in [-1, 2]$ for all maturities. Using the sequence of portfolio weights, we compute average utility, or certainty equivalent return (CER), for each forecast method using (20). We similarly compute the CER for the EH benchmark prediction in lieu of the predictive models. The CER gain is then the difference between the CER for the predictive models and the CER for the EH benchmark. We annualize the CER gain so that it can be interpreted as the annual portfolio management fee that an investor would be willing to pay to have access to the information in the predictive forecast relative to the EH benchmark.[34] Thus, this setup enables a direct assessment of the economic value associated with the increased forecasting precision.

A.   Certainty equivalent returns

Table 7 reports annualized CER gains for all individual bond predictors (for comparison) relative to the EH in Panel A, for EW in Panel B, and for our dynamic forecast combination strategy relative to the EH and the equal-weighted combination strategy in Panels C and D, respectively. In our main results, we set $\gamma = 10$ as in Eriksen (2017), but show in the Internet Appendix that our results are qualitatively similar for other reasonable values of relative risk aversion, e.g. $\gamma = 5$. In order to evaluate the statistical significance of the CER gains, we follow Eriksen (2017) and Gargano et al. (2019) and conduct a conventional t-test on the mean of the time series of realized utility differences. Standard errors are obtained using a Newey and West (1987) estimator.

[Insert Table 7 About Here]

---

[33]We always use the same variance estimated over the same period as the forecasts for all models so that the optimal portfolio weights only differ because of differences in the excess bond return forecast.

[34]Trading costs are generally small in U.S. Treasury bond markets (Adrian, Fleming, and Vogt, 2017).

Overall, we find little evidence that predictable deviations from the EH can be exploited to generate economic value on average when considering individual methods (Panel A). The exception is LN that generally do well utility-wise, something that contrast the statistical results, yet is consistent with the literature that often find disagreeing economical and statistical results. EW (Panel B), on the other hand, obtains positive CER gains for all maturities, indicating that combination forecasts may improve the economic value.

Panel C considers the CER gains for the dynamically combined forecasts based on PMI, $\mathcal{U}$, and the unconditional benchmark NONE. Consistent with our statistical results, we obtain positive CER gains in all cases that are reliably different from zero. The PMI-based dynamic forecasts deliver positive CER gains between 0.39 and 1.43, which are significantly different from zero at conventional significance levels for all maturities. The $\mathcal{U}$-based dynamic forecasts similarly deliver positive values that are significant on conventional significance levels for all maturity bonds except the 2-year bond. NONE always deliver less economic value than using PMI and $\mathcal{U}$, yet all are positive, close to that of the EW, and the 4- and 5-year bonds deliver statistically significant gains on a five percent level. As such, the overall message is clearly supportive of the notion that taking state-dependencies in bond return predictability into account leads to substantial improvements in forecasting accuracy and that these improvement translates into better investment performance for a mean-variance investor that trades in the U.S. Treasury bond market.

Panel D confirms this conclusion by documenting positive CER gains for the dynamic forecast combination strategies relative to EW. Moreover, those increases are all statistically significant at conventional significance levels except in one case. We stress that this strongly supports the idea that dynamically trimming the set of models prior to combining can substantially improve forecast performance and the resulting economic value. That is, eliminating forecasting methods predicted to perform poorly and only maintaining methods with best, yet indistinguishable conditional predictive ability delivers both statistical as well as economic value.

[Insert Figure 7 About Here]

[Insert Figure 8 About Here]

Figures 7 and 8 plots the cumulative realized utilities for our dynamic forecast combination strategies relative to the EH and the EW, respectively. Overall, we note that utility gains are enjoyed uniformly over the out-of-sample period relative to the EH. This is remarkable as our approach is not designed to capture utility, but statistical predictability.[35] This finding is also important in light of the puzzling contradiction between the statistical and economic evidence typically documented in the literature (Thornton and Valente, 2012, Sarno et al., 2016). That is, our conditional view of bond return predictability resolves the puzzle in a simple manner, adding to the findings in Gargano et al. (2019).

B.   State-dependent utility

Table 8 reports state-dependent CER gains for the individual predictors, EW, and dynamically combined forecasts relative to the EH.

[Insert Table 8 About Here]

Individual predictors generally deliver negative CER gains in low (high) economic activity (uncertainty) states and positive CER gains in high (low) economic activity (uncertainty) states. This is fully consistent with the results from the statistical evaluation and suggests that

---

[35]In Appendix IA.D.5 we report resulting improvements in the Sharpe ratio earned by an active investor exploiting predictive information (summarized by the R^2_{OS}) over the Sharpe ratio earned by a buy-and-hold investor, following Gu, Kelly, and Xiu (2020). The results show that individual predictors almost always lead to Sharpe ratio reductions and if any gain is achieved it is generally small. On the other hand, EW provides notable Sharpe ratio improvements, yet our dynamic forecast combination is superior for all bond maturities. For instance, the buy-and-hold Sharpe ratio of a 2-year bond, which is 0.90 in our out-of-sample period, can be improved with 0.48 (0.59) using PMI ($\mathcal{U}$) as state variable. Improvements for NONE are almost identical to EW.

PMI and $\mathcal{U}$ predict not only statistical performance, but economic value as well. LN, however, stands out by delivering positive CER gains across all states and maturities. We note that these findings contrast those of Sarno et al. (2016) who document positive (negative) economic value relative to the EH during times of high (low) uncertainty. Their setting, however, is different as predictions are derived from affine term structure models and not predictive regressions as here. We end this section by drawing attention to the strong and consistently positive CER gains derived from the dynamically combined forecasts across all states and bond maturities. This highlights the strength of dynamically combining forecasts predicted to perform well and shows that applying a conditional view on bond return predictability is key to resolving the contradicting results for statistical and economic evaluations found in the literature.

VII. A model for time-varying predictability

Our empirical findings imply that (i) bond return predictability is state-dependent, (ii) the EH should hold periodically in some states, but not in others, and (iii) the states should be related to economic activity and/or uncertainty such that the EH holds in low (high) economic activity (uncertainty) states. In this section, we show that these findings can be explained within a non-linear term structure model that allows for state-dependencies in the yield curve. Specifically, we consider a version of the non-linear model presented in Feldhütter et al. (2018) and, in addition, draw upon the theoretical insights of Cujean and Hasler (2017). We focus on a one-factor specification to illustrate the predictions of the model in a simple setting and adopt a notation similar to that of Feldhütter et al. (2018). Appendix IA.E provides further details on the model and the estimation procedure using a square-root unscented Kalman filter.

A. The model

The model is an arbitrage-free dynamic term structure model in which the short rate and market prices of risk are non-linear functions of a Gaussian state variable $X(t)$ with dynamics

$$(22) \qquad dX(t) = \kappa \left(\theta - X(t) \right) dt + \Sigma dW(t),$$

where κ, θ, and Σ are constants, and $W(t)$ is a standard Brownian motion. The model assumes a functional form for the stochastic discount factor (SDF)

$$(23) \qquad M(t) = M_0(t) \left(1 + \gamma e^{-\beta X(t)} \right) = M_0(t) + M_1(t),$$

where $M_1(t) = M_0(t)\gamma e^{-\beta X(t)}$, $\gamma > 0$ and β are constants, and $M_0(t)$ is a strictly positive stochastic process that follows the dynamics

$$(24) \qquad dM_0(t) = -r_0(t) M_0(t) dt - \Lambda_0(t) M_0(t) dW(t),$$

where $r_0(t)$ is the short (risk-free) rate, and $\Lambda_0(t)$ the market price of risk. $M_0(t)$ and $M_1(t)$ can be viewed as the SDFs corresponding to two artificial economies: economy 0 and economy 1. We interpret these artificial economies as corresponding to different states of the world. The short rate and market price of risk in the two states are given by

$$(25) \qquad r_i(t) = \rho_{i,0} + \rho_{i,X} X(t), \qquad i = 1,2$$

$$(26) \qquad \Lambda_i(t) = \lambda_{i,0} + \lambda_{i,X} X(t), \qquad i = 1,2$$

$$(27) \qquad \lambda_{1,0} = \lambda_{0,0} + \Sigma\beta,$$

$$(28) \qquad \lambda_{1,X} = \lambda_{0,X},$$

$$(29) \qquad \rho_{1,0} = \rho_{0,0} + \beta\kappa\theta - \frac{1}{2}\beta^2\Sigma^2,$$

$$(30) \qquad \rho_{1,X} = \rho_{0,X} - \kappa\beta - \lambda_{0,X}\Sigma\beta,$$

where all parameters are scalars due to one-dimensionality of $X(t)$. The price of a zero-coupon bond is then a linear combination of the prices in two artificial economies

$$(31) \qquad P(t,T) = s(t)P_0(t,T) + (1 - s(t))P_1(t,T),$$

where $P_i(t,T)$ denotes the zero-coupon bond price in the ith economy and the stochastic weighting between the two states is determined by $s(t) = \frac{1}{1+\gamma e^{-\beta X(t)}} \in [0,1]$. If either $\gamma = 0$ or $\beta = 0$, then the model collapses to a Gaussian term structure model. Provided that the short rate is affine in economy 0, the short rate is also affine in economy 1, implying that $P_i(t,T)$ is exponential affine for both $i = 0, 1$. From Duffie and Kan (1996), we then have that $P_i(t,T) = e^{A_i(\tau) + B_i(\tau)X(t)}$ for $i = 0, 1$. The expected log excess return of buying a zero-coupon bond with k periods to maturity and a holding period of τ in the ith economy is

$$
\begin{aligned}
\mathbb{E}_t\left[rx_{t+\tau}^{(k)}\right] = & A_i(k - \tau) + B_i(k - \tau)\,\mathbb{E}_t[X(t + \tau)] \\
& - A_i(k) - B_i(k)X(t) + A_i(\tau) + B_i(\tau)X(t),
\end{aligned}
$$

(32)

where $A_i(T - t)$ and $B_i(T - t)$ are the solutions to some ordinary differential equations. If the EH holds in economy i, then (32) must be a constant, and, hence, independent of $X(t)$. This restriction is satisfied if $B_i(t) = 0$ for all t and implies, in the ith economy, that the short rate $r_i(t) = \rho_{i,0} + \rho_{i,X}X(t)$ must be independent of $X(t)$ and equal to a constant, i.e. $\rho_{i,X} = 0$.

B. Estimating the model

In estimating the model, we follow Sarno et al. (2016) and base our estimates on both zero-coupon yields and excess holding period returns. More specifically, we include zero-coupon yields with $k = \{12, 24, 36, 48, 60\}$ months to maturity as well as monthly excess returns for bonds with a maturity of $k = \{24, 36, 48, 60\}$ months. The parameters are estimated over the

out-of-sample period (i.e. from January 1990 to December 2018) and we use a square-root unscented Kalman filter. To ensure identification of the parameters when $s(t)$ is close to zero or one, $\gamma = 0$, or $\beta = 0$, we impose the normalization proposed in Dai and Singleton (2000), as do Feldhütter et al. (2018).

C.   Results

The implications described above translate into the following three testable hypotheses:

$\mathbb{H}^{(1)}$: State-dependency requires that both γ and β must be different from zero and that γ is not too large,

$\mathbb{H}^{(2)}$: The EH holding periodically requires that either $\rho_{0,X}$ or $\rho_{1,X}$ is equal to zero,

$\mathbb{H}^{(3)}$: The state weight $s(t)$ must be related to $\mathcal{U}$ and/or PMI and the EH should hold in low (high) PMI ($\mathcal{U}$) states.

[Insert Table 9 About Here]

Panel A of Table 9 presents parameter estimates and p-values from an asymptotic t-test for γ, β, and $\rho_{0,X}$. First, both γ and β are, consistent with $\mathbb{H}^{(1)}$, statistically significant. Hence, nonlinearity is important for explaining yields and excess returns and state-dependency is present.[36] Second, $\rho_{0,X}$ is economically and statistically insignificant and numerically close to zero, implying that the model identifies economy 0 as corresponding to a state of the world in which the EH holds. Moreover, since the estimated $\beta \neq 0$, the short rate in economy 1 is strongly related to $X(t)$ and we conclude that the EH is violated in economy 1. That is, the two economies are distinctly different and consistent with the hypothesis $\mathbb{H}^{(2)}$ that the EH holds periodically.

_____

[36]In unreported results, we perform two likelihood-ratio tests to test for $\gamma = 0$ and $\gamma = 100$. These two hypotheses correspond to $s(t)$ being equal to zero or one for all periods, respectively. We reject both hypotheses and conclude that the two tests confirm the importance of state-dependencies.

The final testable hypothesis from the model is whether the weighting between states $s(t)$, i.e. between economy 0 (EH) and economy 1, is related to economic activity and uncertainty (our conditioning variables). Figure 9 plots $s(t)$ against NBER recession dates and 20% and 80% percentiles of PMI and $\mathcal{U}$, respectively. Increases (decreases) in $s(t)$ correspond to larger weights on economy 0 (economy 1). We highlight three observations: (i) $s(t)$ increases during recessions, (ii) $s(t)$ increases (decreases) during periods with low (high) values of PMI, and (iii) $s(t)$ increases (decreases) during periods with high (low) values of $\mathcal{U}$. This is fully consistent with our empirical results and $\mathbb{H}^{(3)}$. To further quantify this relation, Panel B of Table 9 reports correlations between changes in $s(t)$ and our conditioning variables with p-values for the null of zero correlation. The negative (positive) sign with PMI ($\mathcal{U}$) and magnitude of the correlation coefficients are in line with the figures and our results and both are significantly different from zero at a 5% significance level.

These findings suggest that zero-coupon bond prices (and excess returns) are mainly set in a way consistent with the EH in states of low (high) economic activity (uncertainty), but vary with $X(t)$ in the others. This implies state-dependent predictability of bond excess returns and demonstrates that our empirical findings regarding state-dependencies in bond return predictability are consistent with the predictions of a non-linear term structure model that allows for state-dependencies in yields.

Combining the model with the theoretical insight from Cujean and Hasler (2017) provides a natural explanation for the observed state-dependency of predictability. The functional form of the SDF is similar to that considered in Xiong and Yan (2010), which implies that the model can be interpreted as a reduced-form consumption-based asset pricing model in which two agents with different beliefs co-exist and agree to disagree about the stance of the economy (consumption growth rate). Under this view, $e^{-\beta X(t)}$ represents the Radon-Nikodym derivative between the probability measure associated with future consumption growth applied by the two agents. If the Radon-Nikodym derivative between two agents is

(different from) 1 they have (un)equal expectations. A model-implied estimate of overall economic disagreement may then be represented as

$$\hat{\eta}(t) = |e^{-\hat{\beta}\hat{X}(t)} - 1| \tag{33}$$

where $\hat{\eta}(t) = 0$ corresponds to no disagreement and $\hat{\eta}(t) > 0$ whenever the two agents' expectations differ. The larger the $\hat{\eta}(t)$, the larger the amount of disagreement. Figure 9 depicts the dynamics of estimated $\eta(t)$ and its correlations with PMI and $\mathcal{U}$ are reported in Table 9. Disagreement is negatively (positively) correlated with PMI ($\mathcal{U}$) and is generally increasing in bad states and remains elevated or increases during normal times. Importantly, at the onset of good times (high (low) economic activity (uncertainty)) disagreement is at its highest and then starts tapering off. Cujean and Hasler (2017) theorize in a two-agent economy that predictability is generated in exactly those periods where disagreement is at its highest and subsequently decays, with predictability persisting during a short period following the spikes. This aligns perfectly with our high (low) economic activity (uncertainty) states and how they relate to the estimated $\eta(t)$ in our model, thus providing a coherent explanation of our empirical findings of state-dependent predictability.

VIII. Concluding remarks

We study predictable state-dependencies in bond return predictability and provide empirical evidence that bond return predictability is state-dependent and closely related to economic activity and macroeconomic uncertainty. We show that bond risk premia are predictable in times of high (low) economic activity (uncertainty) states identified using the Purchasing Managers' Index (PMI) and the uncertainty index proposed in Jurado et al. (2015), whereas the EH implication of constant risk premia (no-predictability) provides a reasonable anchor in low (high) economic activity (uncertainty) states. A dynamic forecast combination strategy that averages across forecasting methods predicted to do well delivers substantially more

informative forecasts than both its individual constituents and a simple, static equal-weighted forecast combination scheme. The conclusion hold both across standard statistical evaluation metrics and when considering the economic value to a mean-variance investor. We provide evidence that the improved forecast performance originates from the state variables ability to correctly predict periods in which individual predictors are likely to perform well. Dynamically combined forecasts are strongly countercyclical in contrast to most individual predictors' forecasts and the static (equal-weighted) and dynamic unconditional benchmarks. Last, we show that these findings can be explained within a non-linear term structure model that allows for state-dependencies in yields. The model allows for an assessment of time-varying differences in beliefs among agents, which through the theoretical explanation of Cujean and Hasler (2017) explains the generation of state-dependent predictability documented in this paper.

We end by emphasizing that our econometric methodology of conditional predictive ability is not confined to studies of the Treasury bond market, but may find many and diverse applications across the fields of economics and finance. For example, it would be natural to study the conditional predictive ability of the Goyal and Welch (2008) set of predictors in a multivariate setting as a complement to the large literature on their unconditional performance and relation to the business cycle (Henkel et al., 2011, Dangl and Halling, 2012). Similarly, the approach is likely to be useful in evaluating inflation predictability and identifying periods in which variables such as unemployment rates provides useful information. Finally, we also envision its use in comparing professional forecasters and, in particular, to determine if some forecasters are better than others conditional on being in a certain state. We leave these considerations for future research.

References

Adämmer, P. and R. A. Schüssler (2020). Measuring the equity premium: Mind the news! *Review of Finance 43*(1), 1593–1636.

Adrian, T., M. Fleming, and E. Vogt (2017). An index of Treasury market liquidity: 1991-2017. Federal Reserve Bank of New York Staff Reports, no. 827.

Aiolfi, M., C. Capistrán, and A. Timmermann (2011). Forecast combination. In M. P. Clements and D. F. Hendry (Eds.), *Oxford Handbook of Economic Forecasting*, Chapter 12, pp. 355–388. Oxford University Press.

Aiolfi, M. and C. A. Favero (2005). Model uncertainty, thick modelling and the predictability of stock returns. *Journal of Forecasting 24*(4), 233–254.

Aiolfi, M. and A. Timmermann (2006). Persistence in forecasting performance and conditional combination strategies. *Journal of Econometrics 135*(1-2), 31–53.

Andreasen, M. M., T. Engsted, S. V. Møller, and M. Sander (2020). The yield spread and bond return predictability in expansion and recessions. *Review of Financial Studies*. Forthcoming.

Andrews, D. W. K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica 59*(3), 817–858.

Ang, A. and G. Bekaert (2002). Regime switches in interest rates. *Journal of Business & Economic Statistics 20*(2), 163–182.

Baker, S. R., N. Bloom, and S. J. Davis (2016). Measuring economic policy uncertainty. *Quarterly Journal of Economics 131*(4), 1593–1636.

Bali, T. G., S. J. Brown, and Y. Tang (2017). Is economic uncertainty priced in the cross-section of stock returns? *Journal of Financial Economics 126*, 471–489.

Bates, J. M. and C. W. J. Granger (1969). The combination of forecasts. *Operational Research Quarterly 20*(4), 451–468.

Bauer, M. D. and J. D. Hamilton (2018). Robust bond risk premia. *Review of Financial Studies 31*(2), 399–448.

Bekaert, G., E. Engstrom, and Y. Xing (2009). Risk, uncertainty, and asset prices. *Journal of Financial Economics 91*, 59–82.

Bekaert, G., E. C. Engstrom, and N. R. Xu (2019). The time variation in risk appetite and uncertainty. Working paper, Columbia University.

Berardi, A., M. Markovich, A. Plazzi, and A. Tamoni (2020). Mind the (covergence) gap: Bond predictability strikes back! Working paper, Swiss Finance Institute Research Paper Series No. 19-52.

Berge, T. J. and Ò. Jordà (2011). Evaluating the classification of economic activity into recessions and expansion. *American Economic Journal: Macroeconomics 3*, 246–277.

Bianchi, D., M. Büchner, and A. Tamoni (2020). Bond risk premia with machine learning. *Review of Financial Studies*, Forthcoming.

Bjørnland, H. C., K. Gerdrup, A. S. Jore, C. Smith, and L. A. Thorsrud (2012). Does forecast combination improve Norges bank inflation forecasts? *Oxford Bulletin of Economics and Statistics 74*(2), 163–179.

Bloom, N. (2009). The impact of uncertainty shocks. *Econometrica 77*(3), 623–685.

Borup, D. and E. C. M. Schütte (2020). Asset pricing with data revisions. Working paper, Aarhus University.

Borup, D. and M. Thyrsgaard (2017). Statistical tests for equal predictive ability across multiple forecasting methods. Working paper, Aarhus University.

Buraschi, A. and A. Jiltsov (2007). Habit formation and macroeconomic models of the term structure of interest rates. *Journal of Finance 62*(6), 3009–3063.

Campbell, J. Y. and J. H. Cochrane (1999). By force of habit: A consumption-based explanation of aggregate stock market behavior. *Journal of political Economy 107*(2), 205–251.

Campbell, J. Y. and R. J. Shiller (1991). Yield spreads and interest rate movements: A bird's eye view. *Review of Economic Studies 58*(3), 495–514.

Campbell, J. Y. and S. B. Thompson (2008). Predicting excess stock returns out of sample: Can anything beat the historical average? *Review of Financial Studies 21*(4), 1509–1531.

Carpenter, J. N., F. Lu, and R. F. Whitelaw (2020). The price and quantity of interest rate risk. Working paper, NYU Stern School of Business.

Chauvet, M. and J. Piger (2008). A comparison of the real-time performance of business cycle dating methods. *Journal of Business & Economic Statistics 26*(1), 42–49.

Christiansen, C., J. N. Eriksen, and S. V. Møller (2014). Forecasting US recessions: The role of sentiment. *Journal of Banking and Finance 49*, 459–468.

Christoffersen, P., C. Dorion, K. Jacobs, and L. Karoui (2014). Nonlinear kalman filtering in affine term structure models. *Management Science 60*(9), 2248–2268.

Cieslak, A. and P. Povala (2015). Expected returns in Treasury bonds. *Review of Financial Studies 28*(10), 2859–2901.

Cieslak, A. and P. Povala (2016). Information in the term structure of yield curve volatility. *The Journal of Finance 71*(3), 1393–1436.

Clark, T. E. and M. McCracken (2013). Advances in forecast evaluation. *In Handbook of Economic Forecasting Volume 2, Elsevier B.V.*, 1107–1201.

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

Clark, T. E. and M. W. McCracken (2012). Reality checks and comparison of nested predictive models. *Journal of Business & Economic Statistics 30*(1), 53–66.

Cochrane, J. H. and M. Piazzesi (2005). Bond risk premia. *American Economic Review 95*(1), 138–160.

Cooper, I. and R. Priestley (2009). Time-varying risk premiums and the output gap. *Review of Financial Studies 22*(7), 2801–2833.

Creal, D. D. and J. C. Wu (2020). Bond risk premia in consumption-based models. Working paper, University of Notre Dame.

Cujean, J. and M. Hasler (2017). Why does return predictability concentrate in bad times? *Journal of Finance 72*(6), 2717–2758.

Dai, Q. and K. J. Singleton (2000). Specification analysis of affine term structure models. *The journal of finance 55*(5), 1943–1978.

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

Diebold, F. X. and R. S. Mariano (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics 13*(3), 134–144.

Diebold, F. X. and P. Pauly (1987). Structural change and the combination of forecasts. *Journal of Forecasting 6*(1), 21–40.

Diebold, F. X. and M. Shin (2019). Machine learning for regularized survey forecast combination: Partially-egalitarian lasso and its derivatives. *International Journal of Forecasting 35*(4), 1679–1691.

Drechsler, I. (2013). Uncertainty, time-varying fear, and asset prices. *Journal of Finance 68*(5), 1843–1889.

Duffie, D. and R. Kan (1996). A yield-factor model of interest rates. *Mathematical finance 6*(4), 379–406.

Eriksen, J. N. (2017). Expected business conditions and bond risk premia. *Journal of Financial and Quantitative Analysis 52*(4), 1667–1703.

Fackler, P. (2000). Moments of affine diffusions. Working Paper, North Carolina State University.

Fama, E. and K. R. French (1997). Industry cost of equity. *Journal of Financial Economics 43*(2), 153–193.

Fama, E. F. (1984). Forward and spot exchange rates. *Journal of Monetary Economics 14*, 319–338.

Fama, E. F. and R. R. Bliss (1987). The information in long-maturity forward rates. *American Economic Review 77*(4), 680–692.

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

Farmer, L. E., L. Schmidt, and A. Timmermann (2019). Pockets of predictability. Working paper.

Feldhütter, P., C. Heyerdahl-Larsen, and P. Illeditsch (2018). Risk premia and volatilities in a nonlinear term structure model. *Review of Finance 22*(1), 337–380.

Filipović, D., M. Larsson, and A. B. Trolle (2017). Linear-rational term structure models. *The Journal of Finance 72*(2), 655–704.

Fulop, A., J. Li, and R. Wan (2020). Real-time Bayesian learning and bond return predictability. *Journal of Econometrics*, Forthcoming.

Gargano, A., D. Pettenuzzo, and A. Timmermann (2019). Bond return predictability: Economic value and links to the macroeconomy. *Management Science 65*(2), 508–540.

Genre, V., G. Kenny, A. Meyler, and A. Timmermann (2013). Combining expert forecasts: Can anything beat the simple average? *International Journal of Forecasting 29*, 108–121.

Ghysels, E., C. Horan, and E. Moench (2018). Forecasting through the rearview mirror: Data revisions and bond return predictability. *Review of Financial Studies 31*(2), 678–714.

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

Giacomini, R. and B. Rossi (2010). Forecast comparisons in unstable environments. *Journal of Applied Econometrics 25*(4), 595–620.

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

Gonçalves, S., M. W. McCracken, and B. Perron (2017). Tests of equal accuracy for nested models with estimated factors. *Journal of Econometrics 198*(2), 231–252.

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

Granger, C. W. J. and M. J. Machina (2006). Forecasting and decision theory. *In Handbook of Economic Forecasting, Elsevier B.V. 1*, 82–98.

Granger, C. W. J. and R. Ramanathan (1984). Improved method of combining forecasts. *Journal of Forecasting 3*(2), 197–204.

Granziera, E. and T. Sekhposyan (2019). Predicting relative forecasting performance: An empirical investigation. *International Journal of Forecasting 35*(4), 1636–1657.

Gray, S. F. (1996). Modeling the conditional distribution of interest rates as a regime-switching process. *Journal of Financial Economics 42*, 27–62.

Gu, S., B. Kelly, and D. Xiu (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies 33*(5), 2223–2273.

Gürkaynak, R. S., B. Sack, and J. H. Wright (2007). The U.S. treasury yield curve: 1961 to the present. *Journal of Monetary Economics 54*(8), 2291–2304.

Hamilton, J. D. (1988). Rational-expectations economic analysis of changes in regimes: An investigation of the term structure of interest rates. *Journal of Economic Dynamics and Control 12*, 385–423.

Hansen, P. R., A. Lunde, and J. M. Nason (2011). The model confidence set. *Econometrica 79*(2), 453–497.

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

Huang, D., F. Jiang, G. Tong, and G. Zhou (2020). Are bond returns predictable with real-time macro data? Working paper.

Hubrich, K. and K. D. West (2010). Forecast evaluation of small nested model sets. *Journal of Applied Econometrics 25*, 574–594.

Jose, V. R. R. and R. L. Winkler (2008). Simple robust averages of forecasts: Some empirical results. *International Journal of Forecasting 24*, 163–169.

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

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

Liew, J. and M. Vassalou (2000). Can book-to-market, size and momentum be risk factors that predict economic growth? *Journal of Financial Economics 57*, 221–245.

Litterman, R. and J. Scheinkman (1991). Common factors affecting bond returns. *Journal of Fixed Income 1*(1), 54–61.

Ludvigson, S. C., S. Ma, and S. Ng (2019). Uncertainty and business cycles: Exogenous impulse or endogenous response? *American Economic Journal: Macroeconomics*, Forthcoming.

Ludvigson, S. C. and S. Ng (2009). Macro factors in bond risk premia. *Review of Financial Studies 22*(12), 5027–5067.

Makridakis, S. and R. L. Winkler (1983). Averages of forecasts: Some empirical results. *Management Science 29*(9), 987–1112.

Mariano, R. S. and D. Preve (2012). Statistical tests for multiple forecast comparison. *Journal of Econometrics 169*(1), 123–130.

Marquering, W. and M. Verbeek (2004). The economic value of predicting stock index returns and volatility. *Journal of Financial and Quantitative Analysis 39*(2), 407–429.

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

McCracken, M. W. and S. Ng (2016). FRED-MD: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics 34*(4), 574–589.

Nelson, C. R. and A. F. Siegel (1987). Parsimonious modeling of yield curves. *Journal of Business 60*(4), 473–489.

Newey, W. K. and K. D. West (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica 55*(3), 703–708.

Paye, B. S. and A. Timmermann (2006). Instability of return prediction models. *Journal of Empirical Finance 13*(3), 274–315.

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

Pettenuzzo, D. and A. Timmermann (2011). Predictability of stock returns and asset allocation under structural breaks. *Journal of Econometrics 164*, 60–78.

Pettenuzzo, D. and A. Timmermann (2017). Forecasting macroeconomic variables under model instability. *Journal of Business & Economic Statistics 35*(2), 183–201.

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

Rossi, B. (2013). Advances in forecasting under instability. In G. Elliott and A. Timmermann (Eds.), *Handbook of Economic Forecasting*, Volume 2, Part B, Chapter 21, pp. 1203–1324. Elsevier.

Samuels, J. D. and R. M. Sekkel (2017). Model confidence sets and forecast combination. *International Journal of Forecasting 33*, 48–60.

Sarno, L., P. Schneider, and C. Wagner (2016). The economic value of predicting bond risk premia. *Journal of Empirical Finance 37*, 247–267.

Schrimpf, A. and Q. Wang (2010). A reappraisal of the leading indicator properties of the yield curve under structural instability. *International Journal of Forecasting 26*(4), 836–857.

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

Stock, J. H. and M. W. Watson (2004). Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting 23*, 405–430.

Svensson, L. E. O. (1994). Estimating and interpreting forward interest rates: Sweden 1992-1994. NBER Working Paper No. 4871.

Thornton, D. L. and G. Valente (2012). Out-of-sample predictions of bond excess returns and forward rates: An asset allocation perspective. *The Review of Financial Studies 25*(10), 3141–3168.

Timmermann, A. (2006). Forecast combination. In G. Elliott, C. W. J. Granger, and A. Timmermann (Eds.), *Handbook of Economic Forecasting*, Volume 1, Chapter 4, pp. 135–196. Elsevier.

Timmermann, A. and Y. Zhu (2017). Monitoring forecast performance. Working paper, Rady School of Management.

Van Der Merwe, R. and E. A. Wan (2001). The square-root unscented kalman filter for state and parameter-estimation. In *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, Volume 6, pp. 3461–3464. IEEE.

Wachter, J. A. (2006). A consumption-based model of the term structure of interest rates. *Journal of Financial Economics 79*(2), 365–399.

Wachter, J. A. and M. Warusawitharana (2009). Predictable returns and asset allocation: Should a skeptical investor time the market? *Journal of Econometrics 148*, 162–178.

West, K. D. (1996). Asymptotic inference about predictive ability. *Econometrica 64*(5), 1067–1084.

West, K. D. (2006). Forecast evaluation. *In Handbook of Economic Forecasting, Elsevier B.V. 1*, 100–134.

White, H. (1994). *Estimation, inference and specification analysis.* New York: Cambridge University Press.

White, H. (2001). *Asymptotic theory for econometricians.* San Diego: Academic Press.

Wooldridge, J. M. and H. White (1988). Some invariance principles and central limit theorems for dependent heterogeneous processes. *Econometric Theory 4*(2), 210–230.

Xiong, W. and H. Yan (2010). Heterogeneous expectations and bond markets. *The Review of Financial Studies 23*(4), 1433–1466.

Zellner, A. (1986). On assessing prior distributions and bayesian regression analysis with g-prior distributions. *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno De Finetti 6*, 233–43.

Table 1: Descriptive statistics

This table presents descriptive statistics for monthly excess bond returns. Panel A reports mean, standard deviation, skewness, kurtosis, Sharpe ratios, and first-order autocorrelation (AC(1)) of bond excess returns for two- to five-year bond maturities. Bond returns are in excess of the implied yield on a one-month Treasury bill. Gross returns do not subtract the one-month implied Treasury bill yield. Monthly bond excess returns are constructed using end-of-month Treasury yield data from Gürkaynak et al. (2007). Panel B reports contemporaneous correlations between the excess bond return series. The sample period is January 1962 to December 2018.

	2-year bond	3-year bond	4-year bond	5-year bond
Panel A: Descriptive statistics				
Mean	1.29	1.60	1.85	2.06
Mean (Gross)	5.73	6.04	6.29	6.50
Std. dev.	2.80	3.92	4.95	5.93
Skewness	0.57	0.25	0.08	0.03
Kurtosis	16.68	11.76	8.58	7.05
Sharpe ratio	0.46	0.41	0.37	0.35
AR(1)	0.17	0.15	0.13	0.12
Panel B: Correlations				
2-year bond	1.00			
3-year bond	0.99	1.00		
4-year bond	0.96	0.99	1.00	
5-year bond	0.93	0.97	0.99	1.00

Table 2: Out-of-sample results

This table reports out-of-sample R^2_{OS} values for various linear predictive models for bond excess return. We consider five different predictors: yield spreads (Campbell and Shiller, 1991), forward spreads (Fama and Bliss, 1987), principal components of yields (Litterman and Scheinkman, 1991), the Cochrane and Piazzesi (2005) forward rate factor, and the Ludvigson and Ng (2009) macroeconomic factor. For each model, we report the out-of-sample R^2 from Campbell and Thompson (2008) and the associated Diebold and Mariano (1995) p-value in parenthesis for the null of no predictability implied by the EH (Panels A, B, and C) and against the static (equal-weighted) forecast combination strategy (Panel D). PMI denotes the Purchasing Managers Index published by the Institute for Supply Management and $\mathcal{U}$ is the macroeconomic uncertainty index from Jurado et al. (2015). The out-of-sample evaluation period runs from January 2000 to December 2018.

	2-year	3-year	4-year	5-year
Panel A: Individual bond predictors against EH				
CS	-2.73	-0.53	0.67	1.38
	(0.70)	(0.56)	(0.40)	(0.27)
FB	-0.02	1.31	1.72	1.78
	(0.50)	(0.29)	(0.22)	(0.23)
PC	-9.86	-7.64	-5.91	-4.83
	(0.92)	(0.93)	(0.92)	(0.90)
CP	-6.63	-5.29	-4.27	-3.43
	(0.96)	(0.96)	(0.94)	(0.90)
LN	-7.61	-0.48	1.93	2.43
	(0.73)	(0.52)	(0.42)	(0.39)
Panel B: Static forecast combination against EH				
EW	6.08	5.28	4.89	4.58
	(0.03)	(0.02)	(0.02)	(0.02)
Panel C: Dynamic forecast combination against EH				
PMI	7.98	5.64	5.11	6.16
	(0.01)	(0.02)	(0.02)	(0.01)
$\mathcal{U}$	9.86	6.77	6.09	4.98
	(0.01)	(0.00)	(0.00)	(0.01)
NONE	6.66	5.31	5.25	4.81
	(0.02)	(0.02)	(0.01)	(0.02)
Panel D: Dynamic forecast combination against EW				
PMI	2.02	0.39	0.23	1.66
	(0.01)	(0.24)	(0.33)	(0.06)
$\mathcal{U}$	4.02	1.58	1.26	0.42
	(0.02)	(0.00)	(0.00)	(0.22)
NONE	0.62	0.04	0.38	0.24
	(0.21)	(0.47)	(0.16)	(0.32)

59

Table 3: Testing for equal (un)conditional predictive ability

This table reports full sample multivariate test statistics for equal (un)conditional predictive ability using three different conditioning variables. PMI refers to the case of $\boldsymbol{h}_t = (1, \mathrm{PMI}_t)'$ that is designed to capture business cycle fluctuations. $\mathcal{U}$ refers to the case of $\boldsymbol{h}_t = (1, \mathcal{U}_t)'$ that is chosen to study the effect of macroeconomic uncertainty. NONE refers to an unconditional version of the tests in which $\boldsymbol{h}_t = 1$ for all t. PMI is the Purchasing Managers' Index and UNC is the macroeconomic uncertainty index of Jurado et al. (2015). p-values are presented in parenthesis. The full sample test period runs from January 1990 to December 2018.

	2-year bond	3-year bond	4-year bond	5-year bond
PMI	31.36	34.65	29.76	26.73
	(0.00)	(0.00)	(0.00)	(0.00)
$\mathcal{U}$	27.03	27.95	26.16	26.22
	(0.00)	(0.00)	(0.00)	(0.00)
NONE	8.07	5.68	5.10	5.77
	(0.15)	(0.34)	(0.40)	(0.33)

Table 4: Inclusion frequencies across states

This table reports the inclusion frequencies of the predictor models in three different states of the world identified using the 20% and 80% quantiles of the Purchasing Managers' Index (PMI). We consider five different predictors: yield spreads (Campbell and Shiller, 1991), forward spreads (Fama and Bliss, 1987), principal components of yields (Litterman and Scheinkman, 1991), the Cochrane and Piazzesi (2005) forward rate factor, and the Ludvigson and Ng (2009) macroeconomic factor. EH denotes the benchmark expectations hypothesis model. The out-of-sample evaluation periods runs from January 2000 to December 2018.

	2-year	3-year	4-year	5-year	2-year	3-year	4-year	5-year
	Panel A: Low activity				Panel D: Low uncertainty			
CS	1.00	1.00	1.00	1.00	0.25	0.64	0.64	0.57
FB	0.94	1.00	0.88	0.76	0.25	0.64	0.64	0.57
PC	0.45	0.36	0.42	0.42	1.00	1.00	0.93	1.00
CP	0.88	0.91	1.00	1.00	0.25	0.57	0.54	0.46
LN	0.73	0.67	0.70	0.64	0.96	1.00	1.00	1.00
EH	0.97	1.00	1.00	1.00	0.18	0.50	0.57	0.46
	Panel B: Normal activity				Panel E: Normal uncertainty			
CS	0.86	0.90	0.94	0.98	0.69	0.88	0.86	0.84
FB	0.90	0.98	0.98	0.95	0.79	0.93	0.88	0.76
PC	0.51	0.41	0.47	0.50	0.43	0.53	0.54	0.58
CP	0.65	0.68	0.68	0.62	0.62	0.71	0.63	0.51
LN	0.71	0.79	0.87	0.86	0.83	0.85	0.90	0.92
EH	0.70	0.84	0.85	0.82	0.63	0.83	0.85	0.75
	Panel C: High activity				Panel F: High uncertainty			
CS	0.58	0.75	0.85	0.93	0.95	1.00	1.00	0.98
FB	0.65	0.80	0.88	0.83	0.95	1.00	0.98	0.95
PC	0.83	0.74	0.78	0.88	0.49	0.51	0.44	0.28
CP	0.53	0.53	0.45	0.08	0.95	1.00	0.98	0.98
LN	1.00	1.00	1.00	0.95	0.72	0.98	0.95	1.00
EH	0.33	0.45	0.45	0.48	1.00	1.00	1.00	0.98

Table 5: Out-of-sample R^2 across states

This table reports out-of-sample R^2_{OS} values for various predictive methods for bond excess return conditional on states identified by the Purchasing Manager's Index (PMI) and the the macroeconomic uncertainty index ($\mathcal{U}$) proposed in Jurado et al. (2015). We consider five different predictors: yield spreads (Campbell and Shiller, 1991), forward spreads (Fama and Bliss, 1987), principal components of yields (Litterman and Scheinkman, 1991), the Cochrane and Piazzesi (2005) forward rate factor, and the Ludvigson and Ng (2009) macroeconomic factor. In addition, we report values for a static equal-weighted forecast (EW), dynamically combined forecasts based on PMI and $\mathcal{U}$, and an unconditional alternative based on past performance (NONE). For each method, we report the out-of-sample R^2 from Campbell and Thompson (2008) relative to the expectations hypothesis. High (low) states are identified using the 80% (20%) quantiles of the time series of PMI and $\mathcal{U}$. The out-of-sample evaluation period runs from January 2000 to December 2018.

	2-year	3-year	4-year	5-year	2-year	3-year	4-year	5-year
	Panel A: Low activity				Panel D: Low uncertainty			
CS	-19.73	-10.56	-6.41	-3.85	14.33	9.87	7.79	6.40
FB	-13.17	-6.09	-3.43	-2.22	15.19	8.57	5.00	3.03
PC	-37.08	-26.84	-20.98	-17.02	24.53	16.39	12.09	10.05
CP	-13.30	-7.27	-3.54	-0.83	-83.09	-42.16	-23.01	-12.28
LN	-23.94	-15.44	-12.03	-11.24	14.05	10.04	7.52	5.73
EW	1.34	2.53	2.74	2.47	18.52	12.96	10.19	8.65
PMI	3.55	1.53	2.28	6.30	21.23	13.55	10.59	10.09
$\mathcal{U}$	9.65	4.93	4.37	2.58	26.34	15.83	13.15	10.27
NONE	4.09	4.45	4.19	3.00	14.83	8.77	9.23	8.33
	Panel B: Normal activity				Panel E: Normal uncertainty			
CS	2.16	1.68	1.95	2.20	-2.64	-0.51	0.61	1.28
FB	4.26	3.25	2.90	2.62	1.09	1.78	2.00	2.00
PC	-5.04	-5.29	-4.19	-3.26	-4.86	-4.46	-3.44	-2.74
CP	-6.09	-6.42	-6.06	-5.61	-1.03	-2.14	-2.56	-2.90
LN	-2.58	3.80	5.78	6.25	-2.15	4.31	6.44	7.16
EW	7.77	6.03	5.41	5.10	7.46	5.17	4.37	3.98
PMI	8.96	6.62	5.71	6.02	9.96	6.39	4.95	4.15
$\mathcal{U}$	9.60	7.42	6.77	5.99	9.10	5.96	4.93	4.16
NONE	8.00	5.72	5.46	5.76	7.49	5.02	4.52	4.23
	Panel C: High activity				Panel F: High uncertainty			
CS	6.86	5.32	5.38	5.78	-4.29	-2.03	-0.67	0.30
FB	4.44	3.79	3.83	4.02	-2.82	-0.55	0.50	0.98
PC	20.26	12.96	9.43	7.36	-19.49	-16.08	-13.96	-12.51
CP	9.32	7.30	6.43	6.00	-7.19	-4.90	-3.26	-1.81
LN	-8.41	-2.57	0.26	2.18	-18.80	-11.59	-8.99	-8.73
EW	10.24	7.96	7.13	6.75	2.66	3.77	4.21	4.16
PMI	14.44	10.28	8.37	6.45	3.73	2.85	3.77	8.25
$\mathcal{U}$	11.49	7.97	6.82	5.74	8.37	5.99	5.95	4.51
NONE	7.09	5.50	6.54	4.67	4.50	5.00	5.30	4.55

Table 6: Correlations between forecasts and economic activity

This table reports correlation coefficients between out-of-sample generated forecasts from individual bond predictors (Panel A), the static (equal-weighted) forecast combination strategy (Panel B), and the dynamic forecast combination strategy (Panel C) and economic activity as measured by the Purchasing Managers' Index (PMI). We report p-values for the null of no correlation in parenthesis. The out-of-sample evaluation period runs from January 2000 to December 2018.

	2-year bond	3-year bond	4-year bond	5-year bond
Panel A: Individual bond predictors				
CS	0.36	0.31	0.27	0.23
	(0.00)	(0.00)	(0.00)	(0.00)
FB	0.27	0.19	0.14	0.09
	(0.00)	(0.00)	(0.01)	(0.08)
PC	0.33	0.36	0.36	0.35
	(0.00)	(0.00)	(0.00)	(0.00)
CP	0.15	0.16	0.16	0.16
	(0.01)	(0.00)	(0.00)	(0.00)
LN	-0.38	-0.38	-0.38	-0.38
	(0.00)	(0.00)	(0.00)	(0.00)
EH	0.07	0.14	0.16	0.17
	(0.17)	(0.01)	(0.00)	(0.00)
Panel B: Static forecast combination				
EW	0.01	0.01	0.00	-0.01
	(0.88)	(0.87)	(0.98)	(0.90)
Panel C: Dynamic forecast combination				
PMI	-0.39	-0.40	-0.39	-0.35
	(0.00)	(0.00)	(0.00)	(0.00)
$\mathcal{U}$	-0.40	-0.36	-0.37	-0.39
	(0.00)	(0.00)	(0.00)	(0.00)
NONE	-0.29	-0.26	-0.27	-0.28
	(0.00)	(0.00)	(0.00)	(0.00)

Table 7: Economic Value

This table reports certainty equivalent return (CER) gains for various linear predictive models for bond excess return. We consider five different predictors: yield spreads (Campbell and Shiller, 1991), forward spreads (Fama and Bliss, 1987), principal components of yields (Litterman and Scheinkman, 1991), the Cochrane and Piazzesi (2005) forward rate factor, and the Ludvigson and Ng (2009) macroeconomic factor. For each model, we report the CER gains relative to the expectations hypothesis (Panels A, B and C) and a static (equal-weighted) forecast combination strategy (Panel D). PMI denotes the Purchasing Managers' Index published by the Institute for Supply Management and $\mathcal{U}$ is the macroeconomic uncertainty index from Jurado et al. (2015). CER gains are based on an investor with mean-variance preferences and a relative risk aversion of $\gamma = 10$. The out-of-sample evaluation period runs from January 2000 to December 2018.

	2-year	3-year	4-year	5-year
	Panel A: Individual bond predictors against EH			
CS	-0.64	-0.35	0.10	0.45
	(0.90)	(0.75)	(0.43)	(0.20)
FB	-0.43	-0.12	0.32	0.58
	(0.84)	(0.62)	(0.24)	(0.17)
PC	-1.65	-1.78	-1.65	-1.44
	(0.98)	(0.96)	(0.93)	(0.89)
CP	-0.66	-0.83	-0.76	-0.48
	(0.96)	(0.95)	(0.87)	(0.73)
LN	0.85	1.75	2.32	2.74
	(0.00)	(0.00)	(0.00)	(0.00)
	Panel B: Static forecast combination against EH			
EW	0.10	0.34	0.86	1.07
	(0.36)	(0.16)	(0.03)	(0.02)
	Panel C: Dynamic forecast combination against EH			
PMI	0.39	0.59	1.05	1.43
	(0.08)	(0.06)	(0.02)	(0.00)
$\mathcal{U}$	0.26	0.60	1.17	1.18
	(0.16)	(0.05)	(0.01)	(0.02)
NONE	0.17	0.33	0.92	1.16
	(0.26)	(0.19)	(0.03)	(0.02)
	Panel D: Dynamic forecast combination against EW			
PMI	0.28	0.25	0.19	0.37
	(0.01)	(0.04)	(0.09)	(0.03)
$\mathcal{U}$	0.16	0.25	0.31	0.12
	(0.06)	(0.02)	(0.01)	(0.27)
NONE	0.06	-0.02	0.06	0.09
	(0.13)	(0.56)	(0.24)	(0.26)

Table 8: CER gains across states

This table reports certainty equivalent return (CER) gains for various predictive methods for bond excess return conditional on states identified by the Purchasing Manager's Index (PMI) and the the macroeconomic uncertainty index ($\mathcal{U}$) proposed in Jurado et al. (2015). We consider five different predictors: yield spreads (Campbell and Shiller, 1991), forward spreads (Fama and Bliss, 1987), principal components of yields (Litterman and Scheinkman, 1991), the Cochrane and Piazzesi (2005) forward rate factor, and the Ludvigson and Ng (2009) macroeconomic factor. In addition, we report values for a static equal-weighted forecast (EW), dynamically combined forecasts based on PMI and $\mathcal{U}$, and an unconditional alternative based on past performance (NONE). For each method, we report the CER gain relative to the expectations hypothesis. High (low) states are identified using the 80% (20%) quantiles of the time series of PMI and $\mathcal{U}$. CER gains are based on an investor with mean-variance preferences and a relative risk aversion of $\gamma = 10$. The out-of-sample evaluation period runs from January 2000 to December 2018.

	2-year	3-year	4-year	5-year	2-year	3-year	4-year	5-year
	Panel A: Low activity				Panel D: Low uncertainty			
CS	-3.22	-2.41	-2.04	-2.05	0.02	0.08	1.29	1.68
FB	-2.44	-1.82	-1.83	-2.70	0.01	0.16	0.98	0.86
PC	-7.17	-7.06	-6.92	-6.77	0.02	0.94	2.28	2.76
CP	-2.42	-1.85	-0.91	-0.05	0.08	0.78	2.13	2.70
LN	1.22	2.47	2.05	0.25	0.02	0.41	1.30	1.55
EW	-0.50	0.46	0.88	0.62	-0.01	0.37	1.39	2.10
PMI	0.11	0.58	1.42	1.09	0.05	0.37	1.60	2.26
$\mathcal{U}$	0.07	1.18	1.49	1.14	0.37	0.47	1.93	1.92
NONE	0.21	0.92	1.28	1.25	0.02	0.17	1.24	1.96
	Panel B: Normal activity				Panel E: Normal uncertainty			
CS	-0.19	0.01	0.32	0.48	-0.36	-0.21	0.03	0.24
FB	-0.10	0.04	0.13	-0.01	-0.32	-0.21	0.01	0.07
PC	-1.00	-1.20	-0.96	-0.72	-0.66	-0.72	-0.58	-0.49
CP	-0.35	-0.78	-1.03	-1.02	-0.58	-0.80	-0.68	-0.69
LN	0.82	1.32	1.50	1.78	0.90	1.47	1.72	1.94
EW	0.18	0.21	0.64	0.97	0.04	0.24	0.66	0.92
PMI	0.26	0.31	0.61	1.13	0.38	0.60	0.89	1.06
$\mathcal{U}$	0.26	0.38	0.96	1.04	0.11	0.41	0.84	1.00
NONE	0.22	0.21	0.66	0.96	0.15	0.26	0.63	0.92
	Panel C: High activity				Panel F: High uncertainty			
CS	0.51	0.57	1.43	1.82	-1.41	-0.57	-0.06	-0.18
FB	0.25	0.23	1.17	1.38	-0.82	-0.38	-0.55	-1.69
PC	1.75	2.44	2.84	2.56	-5.19	-5.62	-5.67	-5.60
CP	0.34	0.80	1.61	1.86	-0.74	-1.14	-1.98	-1.42
LN	0.59	1.34	1.99	1.88	1.16	2.30	1.71	0.29
EW	0.31	0.81	1.73	1.85	0.41	0.71	1.23	0.95
PMI	1.17	1.74	2.56	2.17	0.61	0.69	1.28	1.52
$\mathcal{U}$	0.45	0.97	1.74	1.72	0.76	1.37	1.86	1.29
NONE	0.33	0.49	1.29	1.72	0.70	0.85	1.42	1.34

Table 9: Non-linear term structure model

This table reports parameter estimates and test for relevant restrictions for the non-linear one-factor model of Feldhütter et al. (2018). Panel A reports relevant parameters with associated p-values from a t-test with asymptotic standard errors. Panel B provides correlations between the stochastic weighting $s(t)$ and disagreement $\eta(t)$ with and each our conditioning variables (PMI and $\mathcal{U}$) with corresponding p-values for the null of no correlation in parentheses. All estimates are based on yield and excess return data using our out-of-sample observations (i.e. January 1990 to December 2018). The bond maturities are ranging between one and five years for the yields while we consider the excess return of maturities between two and five years with a holding period of one month. The model is estimated using a square-root unscented Kalman filter.

Panel A: Parameters		Panel B: Correlations		
			PMI	$\mathcal{U}_t$
$\rho_{0,X}$	$1.01e^{-06}$			
	(1.00)	$\Delta s(t)$	-0.15	0.13
γ	3.14		(0.01)	(0.01)
	(0.00)	$\Delta \eta(t)$	-0.17	0.11
β	-0.14		(0.00)	(0.04)
	(0.00)			

Figure 1: Bond excess returns

This figure plots times series of monthly bond excess returns (in percentage) for Treasury bonds with maturities ranging from two to five years. Shaded areas represent NBER recession dates. Monthly bond returns are in excess of the implied yield on a one-month Treasury bill rate. Yield data are end-of-month and have been obtained from Gürkaynak et al. (2007) over the period January 1962 to December 2018.
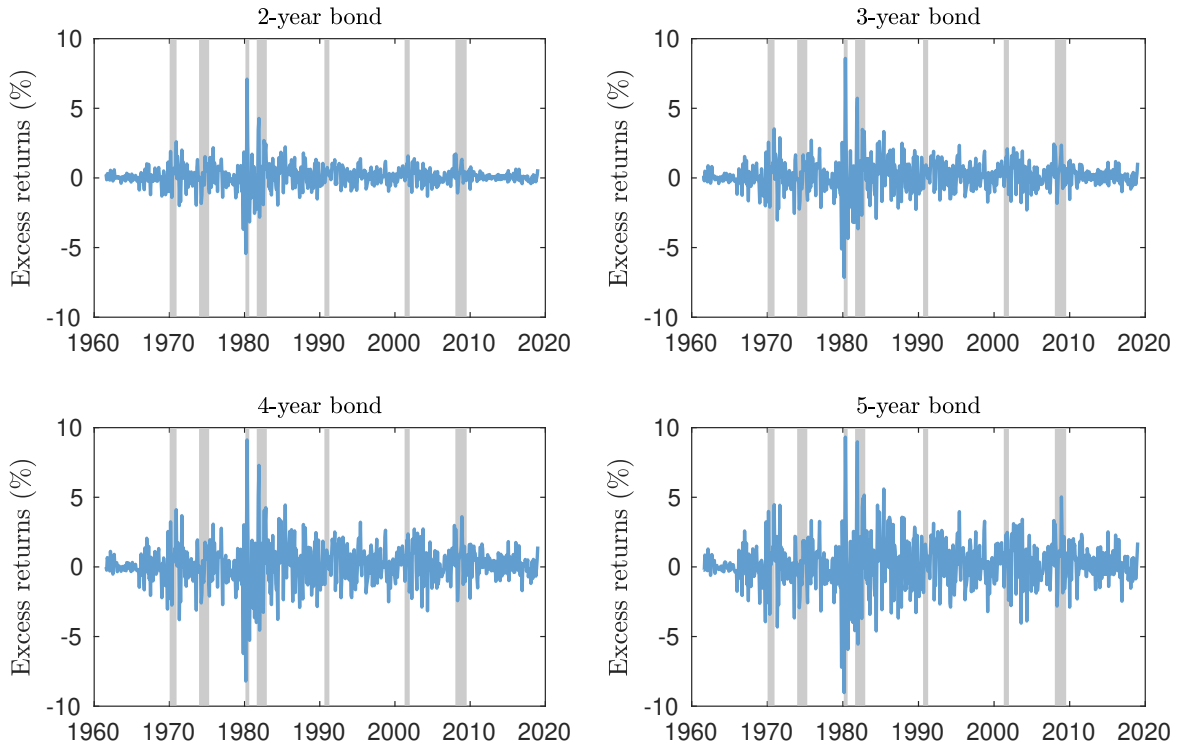
Figure 2: Conditioning variables

This figure shows times series of the Purchasing managers' index (PMI) published by the Institute for Supply Management and the macroeconomic uncertainty ($\mathcal{U}$) index from Jurado et al. (2015). Green (yellow) shaded ares represent periods of (high) low activity and uncertainty, respectively, where high (low) episodes are identified using the 80% (20%) quantiles of their time series. The sample period covers January 1962 to December 2018.

Figure 3: Relative forecasting performance

This figure plots the recursively updated cumulative difference in the squared prediction errors from the EH benchmark model and the ith predictor model over the out-of-sample evaluation period. We consider five different predictors: yield spreads (Campbell and Shiller, 1991), forwa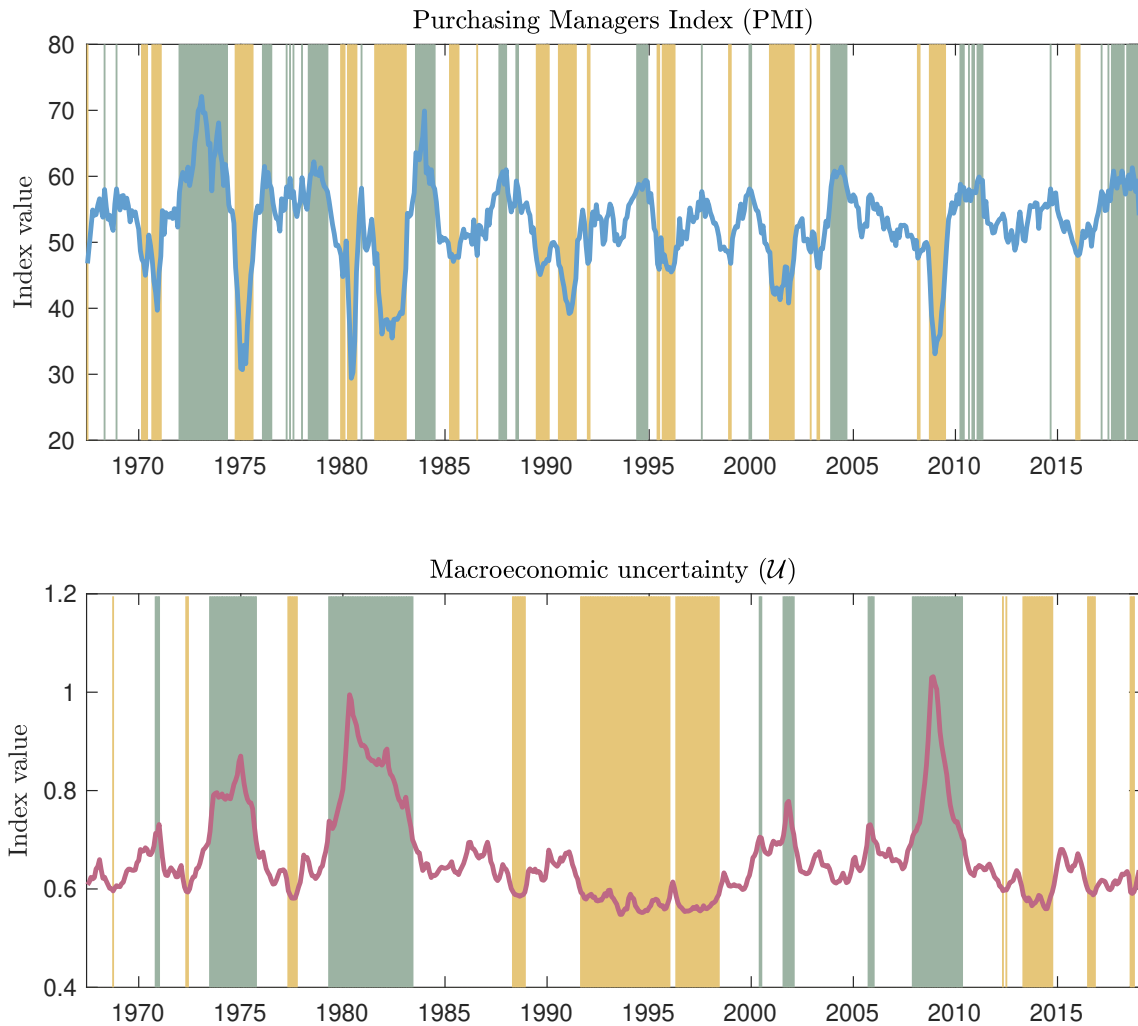rd spreads (Fama and Bliss, 1987), principal components of yields (Litterman and Scheinkman, 1991), the Cochrane and Piazzesi (2005) forward rate factor, and the Ludvigson and Ng (2009) macroeconomic factor. We also consider a simple equal-weighted combination of the individual forecasts. A positive (negative) slope indicates that the predictive model delivers more (less) accurate forecasts than the EH benchmark. Green (yellow) shaded ares represent periods of high (low) activity and uncertainty, respectively, where activity is measured using the Purchasing Managers' Index (PMI) (left column) and uncertainty (right column) is the index developed by Jurado et al. (2015). High (low) episodes are identified using the 80% (20%) quantiles of their time series. White areas are normal times. The out-of-sample evaluation period runs from January 2000 to December 2018.

Figure 4: Dynamic forecast combinations

This figure plots the recursively updated cumulative difference in the squared prediction errors from the EH benchmark model and the dynamic forecast combination forecast for each of the tree conditioning cases. We consider the Purchasing Managers' Index (PMI) and the macroeconomic uncertainty index ($\mathcal{U}$) from Jurado et al. (2015) as our conditioning variables along with an unconditional version labeled NONE. A positive (negative) slope indicates that the dynamic forecast combination delivers more (less) accurate forecasts than the EH benchmark. Green (yellow) shaded ares represent periods of high (low) activity and uncertainty, respectively, where high (low) episodes are identified using the 80% (20%) quantiles of their time series. White areas are normal times. The out-of-sample evaluation periods runs from January 2000 to December 2018.

Figure 5: Dynamic versus static forecast combination

This figure plots the recursively updated cumulative difference in the squared prediction errors from a static equal-weighted forecast combination benchmark and the dynamic forecast combination forecast for each of the tree conditioning cases. We consider the Purchasing Managers' Index (PMI) and the macroeconomic uncertainty index ($\mathcal{U}$) from Jurado et al. (2015) as our conditioning variables along with an unconditional version labeled NONE. A positive (negative) slope indicates that the dynamic forecast combination delivers more (less) accurate forecasts than the static equal-weighted forecast combination benchmark. Green (yellow) shaded ares represent periods of high (low) activity and uncertainty, respectively, where high (low) episodes are identified using the 80% (20%) quantiles of their time series. White areas are normal times. The out-of-sample evaluation periods runs from January 2000 to December 2018.
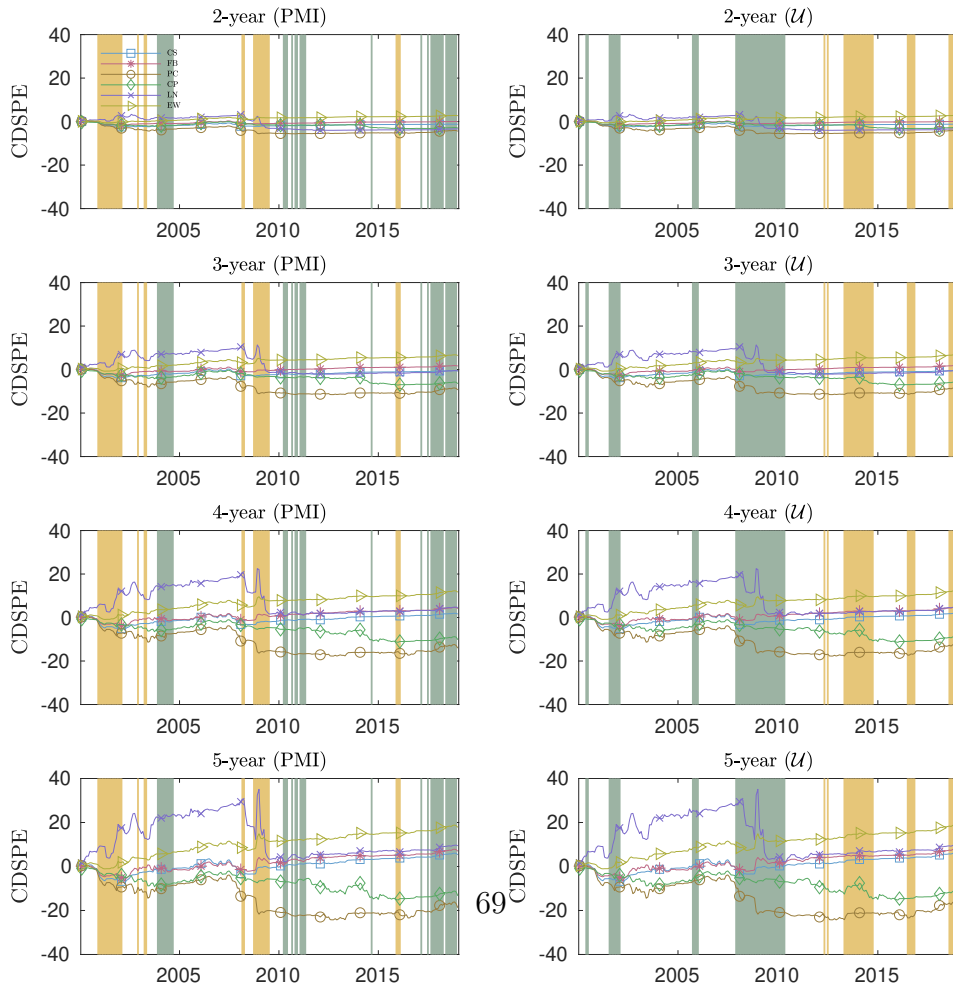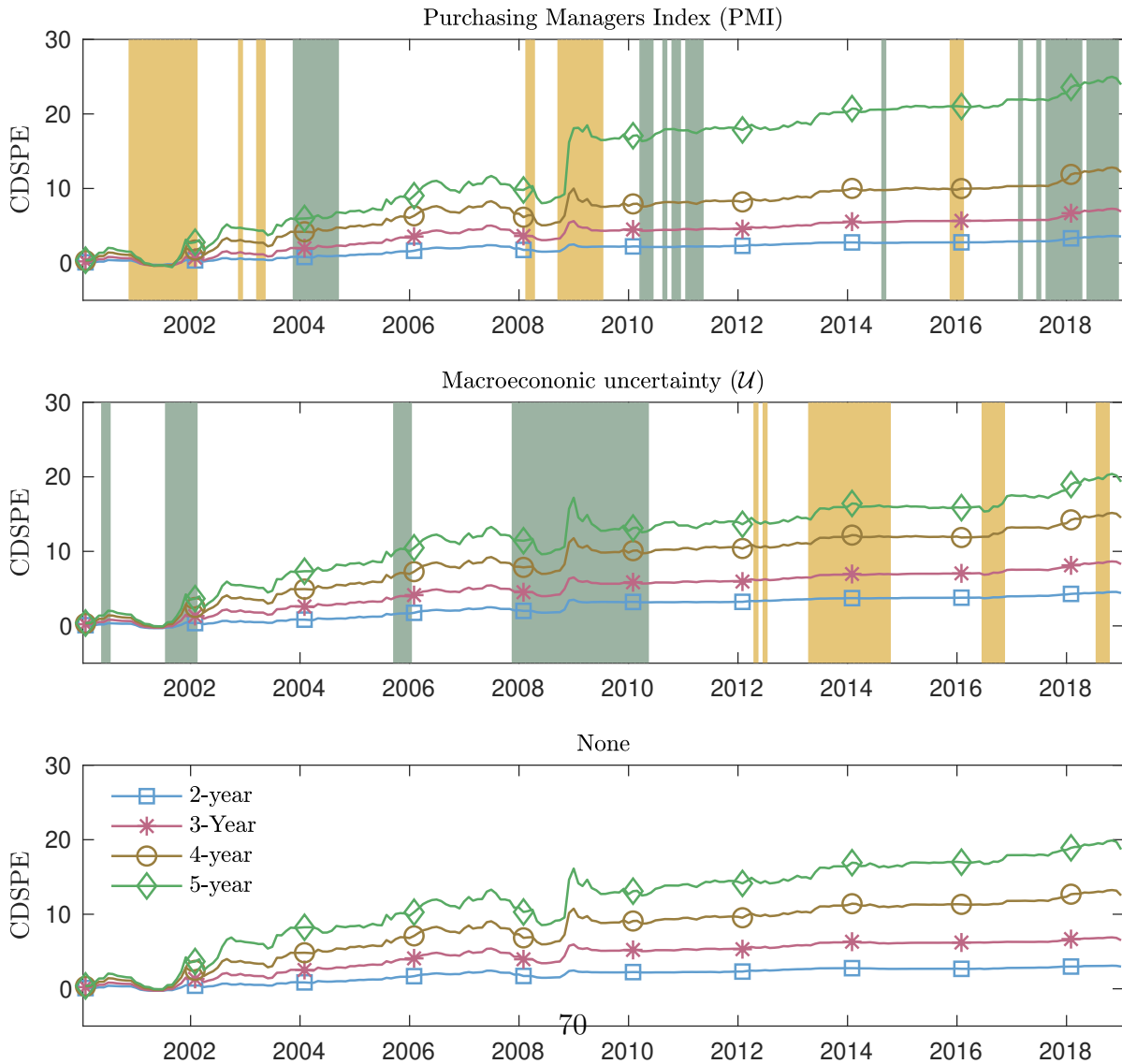

71

Figure 6: Bond risk premia forecasts for dynamic combination strategy

This figure illustrates the time series behavior of bond risk premia forecasts originating from our dynamic forecast combination strategy. Shaded areas represent NBER recession dates. PMI is the Purchasing Managers' Index published by the Institute for Supply Management and $\mathcal{U}$ is the macroeconomic uncertainty index proposed in Jurado et al. (2015). The out-of-sample forecasting periods runs from January 2000 to December 2018.
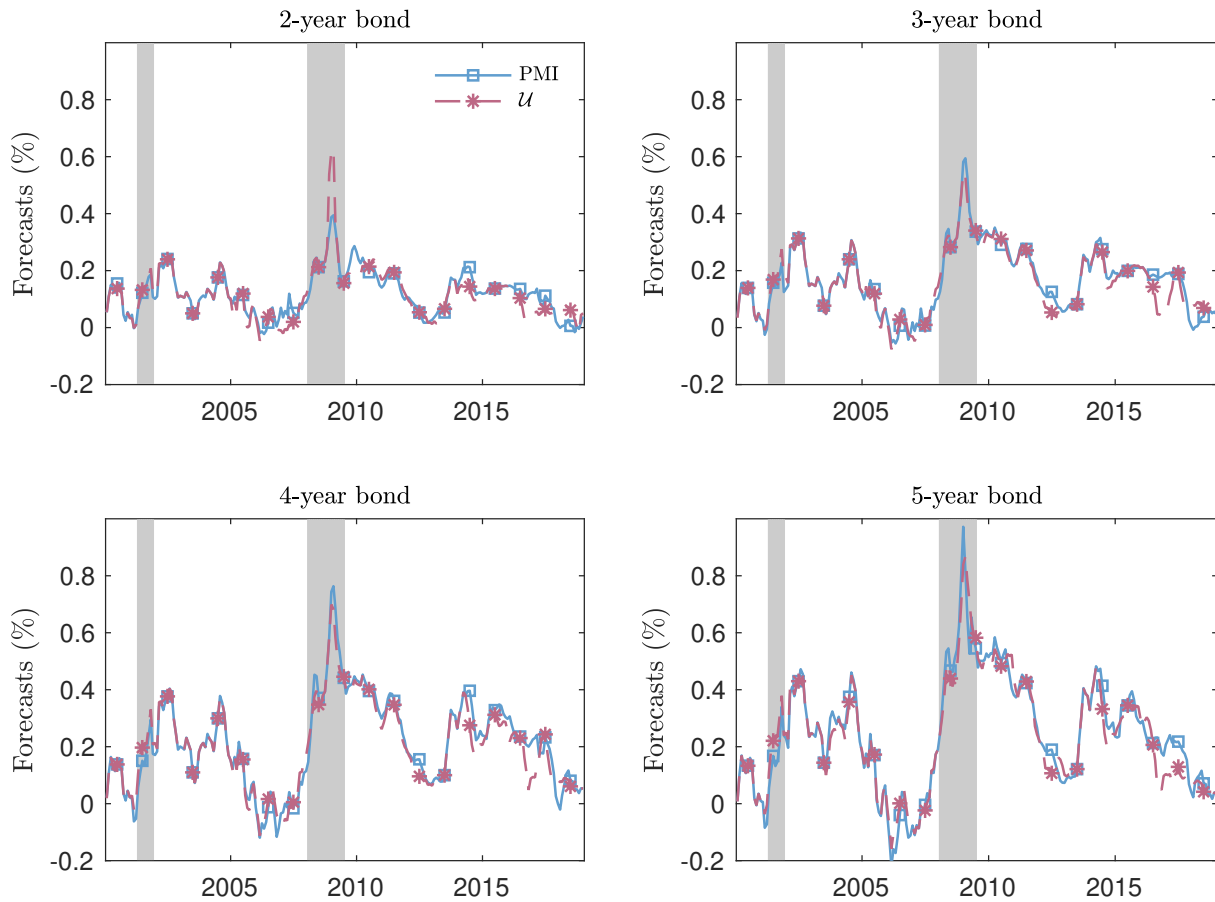
Figure 7: Dynamic forecast combinations: CER gains

This figure plots the recursively updated cumulative difference in realized utility from the dynamic forecast combination forecast for each of the tree conditioning cases and the EH benchmark model. We consider the Purchasing Managers' Index (PMI) and the macroeconomic uncertainty index ($\mathcal{U}$) from Jurado et al. (2015) as our conditioning variables along with an unconditional version labeled NONE. A positive (negative) slope indicates that the dynamic forecast combination delivers more (less) accurate forecasts than the EH benchmark. Green (yellow) shaded ares represent periods of high (low) activity and uncertainty, respectively, where high (low) episodes are identified using the 80% (20%) quantiles of their time series. White areas are normal times. The out-of-sample evaluation periods runs from January 2000 to December 2018.
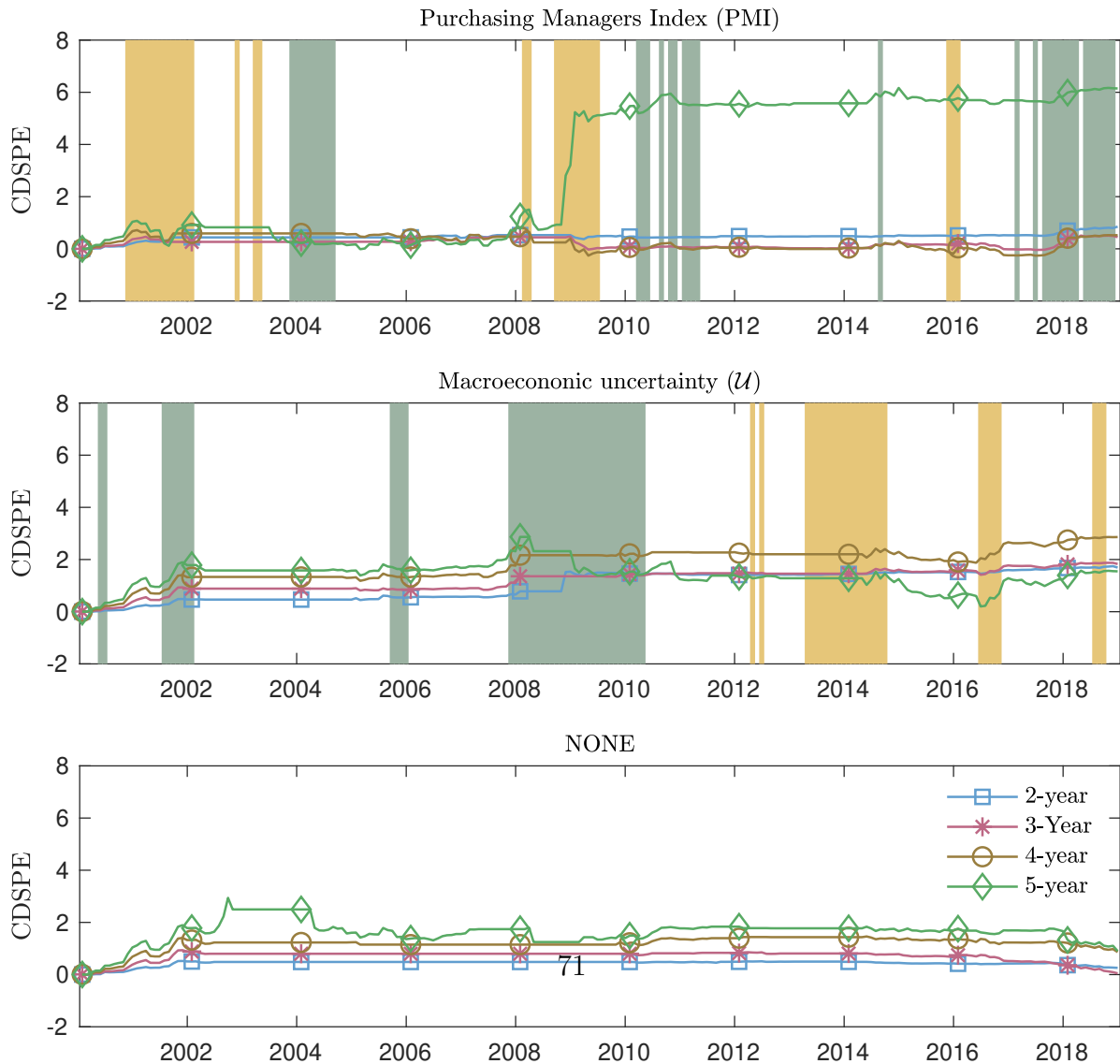
Figure 8: Dynamic versus static forecast combination: CER gains

This figure plots the recursively updated cumulative difference in the squared prediction errors from the dynamic forecast combination forecast for each of the tree conditioning cases and a static equal-weighted forecast combination benchmark. We consider the Purchasing Managers' Index (PMI) and the macroeconomic uncertainty index ($\mathcal{U}$) from Jurado et al. (2015) as our conditioning variables along with an unconditional version labeled NONE. A positive (negative) slope indicates that the dynamic forecast combination delivers more (less) accurate forecasts than the static equal-weighted forecast combination benchmark. Green (yellow) shaded ares represent periods of high (low) activity and uncertainty, respectively, where high (low) episodes are identified using the 80% (20%) quantiles of their time series. White areas are normal times. The out-of-sample evaluation periods runs from January 2000 to December 2018.

Figure 9: Stochastic weights on economy 0 and disagreement

The figure illustrates the time series of the estimated stochastic weights $s(t)$ on economy 0 as well as the estimated disagreement $\eta(t)$ in the non-linear model term structure model. $s(t) = 1$ corresponds to an economy in which the expectation hypothesis (EH) holds, and $\eta(t) = 0$ ($\eta(t) > 0$) corresponds to an economy without (with) disagreement. Grey shaded areas represent NBER recessions. Green (yellow) shaded areas represent periods of high (low) activity and uncertainty, respectively, where high (low) episodes are identified using the 80% (20%) quantiles of their time series. White areas are normal times. The model is estimated using data from January 1990 to December 2018.
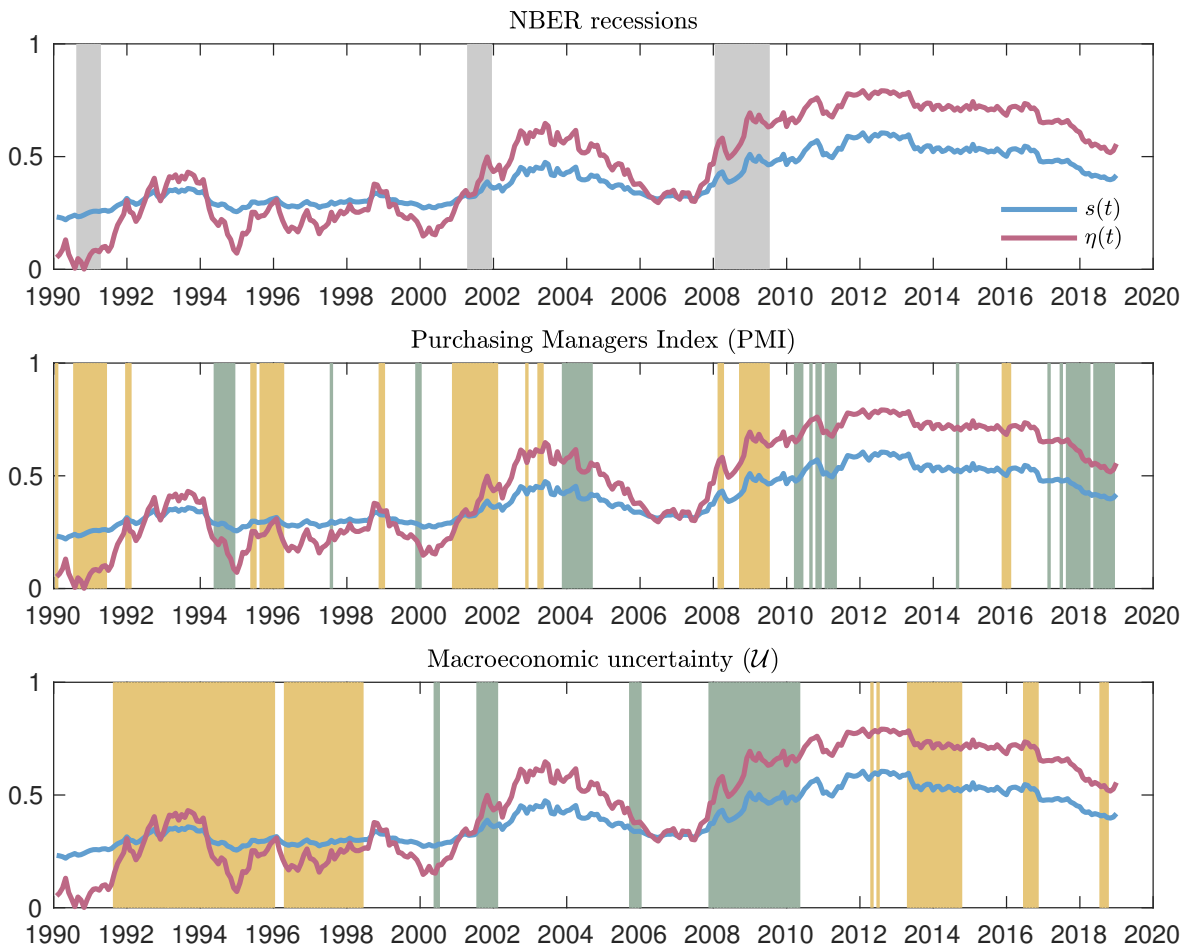
Internet Appendix for

Predicting bond return predictability

(not intended for publication)

IA.A.   Theoretical results, assumptions, and proofs

This section explains the adopted Giacomini and White (2006) assumptions used in Theorem 1 along with its proof. The outline of several of the proofs follows Giacomini and White (2006), making the necessary adjustments to account for the multivariate nature of our tests. We also provide theoretical results with associated proofs for the case of multi-step ahead forecasting, $\tau > 1$, and the unconditional case, $\mathcal{G}_t = \{\emptyset, \Omega\}$.

IA.A.1.   One-step ahead forecasting and Giacomini and White (2006) assumptions

In the one-step ahead case, $\tau = 1$, we impose the following assumptions that are adopted from Giacomini and White (2006).

Assumption 1. $\{\boldsymbol{h}_t\}$ and $\{\boldsymbol{w}_t\}$ are ϕ-mixing with $\phi(t) = O\left(t^{-r/(2r-1)-\iota}\right)$, $r \geq 1$, or α-mixing with $\alpha(t) = O\left(t^{-\frac{r}{r-1}-\iota}\right)$, $r > 1$, for some $\iota > 0$.

Assumption 1 imposes relatively mild restrictions on the dependence structure and heterogeneity of data. We do not impose the stricter and common (covariance) stationarity assumption as used in for instance Diebold and Mariano (1995) and Mariano and Preve (2012). Specifically, data may exhibit arbitrary structural changes, which is a common feature found in many empirical studies within macroeconomic prediction (see e.g. Stock and Watson (2003) and Schrimpf and Wang (2010)), stock return prediction (see e.g. Fama and French (1997) and Paye and Timmermann (2006)), and exchange rate prediction (see e.g. Giacomini and Rossi (2010)), to name a few, and is especially relevant in our context of possible instabilities.

Assumption 2. $\mathbb{E}[|\boldsymbol{d}_{t+1,i}|^{2(r+\delta)}] < \infty$ for some $\delta > 0$, $i = 1, \ldots, qp$, and for all t, where subscript i indicate the ith element of $\boldsymbol{d}_{t+1}$.

Assumption 3. $\boldsymbol{\Sigma}_T \equiv T^{-1} \sum_{t=1}^{T} \mathbb{E}[\boldsymbol{d}_{t+1}\boldsymbol{d}_{t+1}']$ is uniformly positive definite.

Assumptions 2–3 are mainly technical assumptions ensuring (uniformly) bounded moments of data and positive definiteness of the asymptotic variance. Both assumptions are commonly found in the forecast evaluation literature.

IA.A.1.1. Proof of Theorem 1

The proof of part A. and B. adopts the necessary steps in Giacomini and White (2006). We start by proving part A. Let $\boldsymbol{d}_{t+1} = \boldsymbol{h}_t \otimes \Delta \boldsymbol{L}_{+1}$ and write

$$(\text{IA.A.1}) \qquad \boldsymbol{d}_{t+1}\boldsymbol{d}'_{t+1} = g\left(\boldsymbol{h}_t, \boldsymbol{w}_{t+1}, \ldots, \boldsymbol{w}_{t-m}\right)$$

for some measurable function g. Since $m < \infty$, and $\{\boldsymbol{h}_t\}$ and $\{\boldsymbol{w}_t\}$ are mixing of the same size according to Assumption 1, it follows from Theorem 3.49 in White (2001) that $\{\boldsymbol{d}_{t+1}\boldsymbol{d}'_{t+1}\}$ is mixing of the same size as $\{\boldsymbol{h}_t\}$ and $\{\boldsymbol{w}_t\}$.

By Assumption 2, there exists a $\overline{C} \in \mathbb{R}_+$ and $\delta > 0$ such that $\mathbb{E}[|\boldsymbol{d}_{m,t+1,i}|^{2(r+\delta)}] < \overline{C} < \infty$ for $i = 1, \ldots, qp$ and for all t, where subscript i indicates the ith element in $\boldsymbol{d}_{t+1}$. Hence, by the Cauchy-Schwartz inequality, one obtains

$$(\text{IA.A.2}) \qquad \mathbb{E}[|\boldsymbol{d}_{t+1,i}\boldsymbol{d}_{t+1,j}|^{r+\delta}] \leq \mathbb{E}[|\boldsymbol{d}^2_{t+1,i}|^{r+\delta}]^{1/2}\mathbb{E}[|\boldsymbol{d}^2_{t+1,j}|^{r+\delta}]^{1/2} < \overline{C}$$

for $i, j = 1, \ldots, qp$ and for all t. By Corollary 3.48 in White (2001), it then follows that $\widehat{\boldsymbol{\Sigma}}_T - \boldsymbol{\Sigma}_T \xrightarrow{\mathbb{P}} 0$. Furthermore, by Assumption 2, it follows that $\boldsymbol{\Sigma}_T$ is finite and, by Assumption 3, that it is uniformly positive definite.

Next, let $\boldsymbol{\lambda} \in \mathbb{R}^{qp}$ with $\boldsymbol{\lambda}'\boldsymbol{\lambda} = 1$ and consider

$$(\text{IA.A.3}) \qquad \boldsymbol{\lambda}'\boldsymbol{\Sigma}_T^{-1/2}\sqrt{T}\overline{\boldsymbol{d}}_{t+1} = T^{-1/2}\sum_{t=1}^{T-1}\boldsymbol{\lambda}'\boldsymbol{\Sigma}_T^{-1/2}\boldsymbol{d}_{t+1}.$$

Let $\widetilde{\boldsymbol{\lambda}}_i$ denote the ith element of the product $\boldsymbol{\lambda}'\boldsymbol{\Sigma}_T^{-1/2}$, such that $\boldsymbol{\lambda}'\boldsymbol{\Sigma}_T^{-1/2}\boldsymbol{d}_{t+1} = \sum_{i=1}^{qp}\widetilde{\boldsymbol{\lambda}}_i\boldsymbol{d}_{t+1,i}$.

Hence, under the null hypothesis

(IA.A.4)
$$\mathbb{E}[\boldsymbol{\lambda}'\boldsymbol{\Sigma}_T^{-1/2}\boldsymbol{d}_{t+1}|\mathcal{G}_t] = \mathbb{E}\left[\sum_{i=1}^{qp}\widetilde{\boldsymbol{\lambda}}_i d_{t+1,i}|\mathcal{G}_t\right] = \sum_{i=1}^{qp}\widetilde{\boldsymbol{\lambda}}_i\mathbb{E}[d_{t+1,i}|\mathcal{G}_t] = 0,$$

by measurability of $\widetilde{\boldsymbol{\lambda}}_i$, such that the sequence $\{\boldsymbol{\lambda}'\boldsymbol{\Sigma}_T^{-1/2}\boldsymbol{d}_{t+1}, \mathcal{G}_t\}$ is an MDS. The asymptotic variance is

$$\sigma_d^2 = \text{Var}[\boldsymbol{\lambda}'\boldsymbol{\Sigma}_T^{-1/2}\sqrt{T}\overline{\boldsymbol{d}}]$$

$$= \boldsymbol{\lambda}'\boldsymbol{\Sigma}_T^{-1/2}\text{Var}[\sqrt{T}\overline{\boldsymbol{d}}]\boldsymbol{\Sigma}_T^{-1/2}\boldsymbol{\lambda}$$

$$= \boldsymbol{\lambda}'\boldsymbol{\Sigma}_T^{-1/2}\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_T^{-1/2}\boldsymbol{\lambda}$$

(IA.A.5)
$$= 1$$

for sufficiently large T. Furthermore, since $\widehat{\boldsymbol{\Sigma}}_T - \boldsymbol{\Sigma}_T \xrightarrow{\mathbb{P}} 0$ it follows by the Continuous Mapping Theorem that

$$\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\lambda}'\boldsymbol{\Sigma}_T^{-1/2}\boldsymbol{d}_{t+1}'\boldsymbol{d}_{t+1}\boldsymbol{\Sigma}_T^{-1/2}\boldsymbol{\lambda} - \sigma_d^2$$

(IA.A.6)
$$= \boldsymbol{\lambda}'\boldsymbol{\Sigma}_T^{-1/2}\widehat{\boldsymbol{\Sigma}}_T\boldsymbol{\Sigma}_T^{-1/2}\boldsymbol{\lambda} - \boldsymbol{\lambda}'\boldsymbol{\Sigma}_T^{-1/2}\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_T^{-1/2}\boldsymbol{\lambda} \xrightarrow{\mathbb{P}} 0.$$

Lastly, we need to check that $\boldsymbol{\lambda}'\boldsymbol{\Sigma}_T^{-1/2}\boldsymbol{d}_{t+1}$ has absolute $2+\delta$ moment. By Minkowski's inequality and Assumption 2 we obtain

$$\mathbb{E}[|\boldsymbol{\lambda}'\boldsymbol{\Sigma}_T^{-1/2}\boldsymbol{d}_{t+1}|^{2+\delta}] = \mathbb{E}\left[\left|\sum_{i=1}^{qp}\widetilde{\boldsymbol{\lambda}}_i d_{t+1,i}\right|^{2+\delta}\right]$$

(IA.A.7)
$$\leq \left(\sum_{i=1}^{qp}\widetilde{\boldsymbol{\lambda}}_i\mathbb{E}\left[|d_{t+1,i}|^{2+\delta}\right]^{1/(2+\delta)}\right)^{2+\delta} < \infty.$$

Consequently, we can apply the CLT for MDS and deduce that $\boldsymbol{\lambda}'\boldsymbol{\Sigma}_T^{-1/2}\sqrt{T}\overline{\boldsymbol{d}} \xrightarrow{d} N(0,1)$. By

the Cramér-Wold device it then follows that

$$(\text{IA.A.8}) \qquad\qquad \boldsymbol{\Sigma}^{-1/2}\sqrt{T}\overline{\boldsymbol{d}} \xrightarrow{d} N(0, \boldsymbol{I}_{qp}).$$

Since $\widehat{\boldsymbol{\Sigma}}_T - \boldsymbol{\Sigma}_T \xrightarrow{\mathbb{P}} 0$, we deduce that

$$(\text{IA.A.9}) \qquad \sqrt{T}\left(\widehat{\boldsymbol{\Sigma}}_T^{-1/2}\overline{\boldsymbol{d}}\right)'\sqrt{T}\boldsymbol{\Sigma}_T^{-1/2}\overline{\boldsymbol{d}} = T\overline{\boldsymbol{d}}'\widehat{\boldsymbol{\Sigma}}_T^{-1}\overline{\boldsymbol{d}} = S_h \xrightarrow{d} \chi^2(qp),$$

as $T \to \infty$. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ □

We now prove part B. By the same arguments as in the proof for part A., it follows that the sequence $\{\boldsymbol{d}_{t+1}\}$ is mixing of the same size as $\{\boldsymbol{w}_t\}$ and $\{\boldsymbol{h}_t\}$. Furthermore, Assumption 2 ensures that each element of $\boldsymbol{d}_{t+1}$ is bounded uniformly in t such that

$$(\text{IA.A.10}) \qquad\qquad \overline{\boldsymbol{d}} - \mathbb{E}[\overline{\boldsymbol{d}}] \xrightarrow{\mathbb{P}} 0$$

by Corollary 3.48 in White (2001). Under the alternative hypothesis there exists $\eta > 0$ such that $\mathbb{E}[\overline{\boldsymbol{d}}_m']\mathbb{E}[\overline{\boldsymbol{d}}_m] > 2\eta$ for T sufficiently large. It follows that

$$\mathbb{P}[\overline{\boldsymbol{d}}'\overline{\boldsymbol{d}} > \eta] \geq \mathbb{P}[\overline{\boldsymbol{d}}'\overline{\boldsymbol{d}} - \mathbb{E}[\overline{\boldsymbol{d}}']\mathbb{E}[\overline{\boldsymbol{d}}] > -\eta]$$
$$(\text{IA.A.11}) \qquad\qquad \geq \mathbb{P}[|\overline{\boldsymbol{d}}_m'\overline{\boldsymbol{d}} - \mathbb{E}[\overline{\boldsymbol{d}}']\mathbb{E}[\overline{\boldsymbol{d}}]| < \eta] \to 1,$$

where the convergence to unity is due to (IA.A.10). By identical arguments as the proof of part A., $\boldsymbol{d}_{t+1}'\boldsymbol{d}_{t+1}$ is mixing with the same size as $\{\boldsymbol{w}_t\}$ and each element is uniformly bounded in t. Corollary 3.48 in White (2001) can then be applied, and it follows that $\widehat{\boldsymbol{\Sigma}}_T$ is a consistent estimator of $\boldsymbol{\Sigma}_T$. By Assumption 3, $\boldsymbol{\Sigma}_T$ is uniformly positive definite. Let $c \in \mathbb{R}_+$.

It then follows from Theorem 8.13 in White (1994) that

(IA.A.12) $$\mathbb{P}[S_h > c] \to 1, \quad \text{as } T \to \infty.$$

$\square$

Last, we prove part C. Let $\boldsymbol{L}^*_{t+1}$ be an arbitrary permutation of the forecasting losses, i.e. $\boldsymbol{L}^*_{t+1} = \boldsymbol{P}\boldsymbol{L}_{t+1}$, where $\boldsymbol{P}$ is a $(p+1)\times(p+1)$ permutation matrix and $\boldsymbol{L}_{t+1} = (L^1_{t+1}, \ldots, L^{p+1}_{t+1})'$. Define the $p \times (p+1)$ matrix $\boldsymbol{D}$ by

$$\boldsymbol{D} = \begin{bmatrix} 1 & -1 & 0 & \ldots & 0 \\ 0 & 1 & -1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & 1 & -1 \end{bmatrix}$$

such that $\Delta\boldsymbol{L}^*_{t+1} = \boldsymbol{D}\boldsymbol{L}^*_{t+1} = \boldsymbol{D}\boldsymbol{P}\boldsymbol{L}_{t+1}$. In total, the number of permutations of the forecast losses at each point of time t is $(p+1)!$. Mariano and Preve (2012) show that there always exists a nonsingular matrix $\boldsymbol{B}$ of dimension $p \times p$ such that $\boldsymbol{B}\Delta\boldsymbol{L}_{t+1} = \Delta\boldsymbol{L}^*_{t+1}$. Consequently, define the $qp \times qp$ matrix $\boldsymbol{A} = (\boldsymbol{I}_q \otimes \boldsymbol{B})$, where $\boldsymbol{I}_q$ is the $q \times q$ identity matrix. By standard properties of the Kronecker product $\boldsymbol{A}$ is nonsingular, and we have that

(IA.A.13) $\quad \boldsymbol{d}^*_{t+1} = \boldsymbol{h}_t \otimes \Delta\boldsymbol{L}^*_{t+1} = (\boldsymbol{I}_q\boldsymbol{h}_t) \otimes (\boldsymbol{B}\Delta\boldsymbol{L}_{t+1}) = (\boldsymbol{I}_q \otimes \boldsymbol{B})(\boldsymbol{h}_t \otimes \Delta\boldsymbol{L}_{t+1}) = \boldsymbol{A}\boldsymbol{d}_{t+1}.$

Since the null hypothesis implies that the asymptotic variance can be estimated consistently by the sample variance, it follows that

$$\widehat{\boldsymbol{\Sigma}}^*_T \equiv \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{d}^*_{t+1}\boldsymbol{d}^{*'}_{t+1} = \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{A}\boldsymbol{d}_{t+1}\boldsymbol{d}'_{t+1}\boldsymbol{A}' = \boldsymbol{A}\widehat{\boldsymbol{\Sigma}}_T\boldsymbol{A}'.$$

Due to the nonsingularity of $\boldsymbol{A}$ and $\widehat{\boldsymbol{\Sigma}}_T$, it follows that

$$\overline{\boldsymbol{d}}_{t+1}^{*'}(\widehat{\boldsymbol{\Sigma}}_T^*)^{-1}\overline{\boldsymbol{d}}_{t+1}^* = \boldsymbol{d}_{t+1}'\boldsymbol{A}'(\boldsymbol{A}\widehat{\boldsymbol{\Sigma}}_T\boldsymbol{A}')^{-1}\boldsymbol{A}\boldsymbol{d}_{t+1}$$
$$= \boldsymbol{d}_{t+1}'\widehat{\boldsymbol{\Sigma}}_T^{-1}\boldsymbol{d}_{t+1},$$

which shows that the test is invariant to a permutation of the ordering of the forecast losses. □

IA.A.2. Unconditional and multi-step predictive ability tests

In both the unconditional, $G_t = \{\emptyset, \Omega\}$, and multistep conditional case the loss series are no longer martingale difference sequences under the null hypothesis. Thus, the sequence $\{\boldsymbol{h}_t \otimes \Delta\boldsymbol{L}_{t+\tau}\}$ may be serially autocorrelated.[37] In the conditional setting, the null hypothesis imposes a particular structure on the serial correlation, namely that it can be at most order $\tau - 1$. However, in the unconditional case no such restriction exists. Consequently, we can no longer rely on the sample variance under the null for estimating the covariance matrix as was the case in the one-step conditional setting considered in the previous section. Instead, we consider a HAC-type estimator (see, e.g., Newey and West (1987) and Andrews (1991)) with a bandwidth choice guided by the implications of the null hypothesis. The estimator is given by

$$\widetilde{\boldsymbol{\Sigma}}_T = \frac{1}{T}\Bigg[\sum_{t=1}^{T}\boldsymbol{d}_{t+\tau}\boldsymbol{d}_{t+\tau}'$$

(IA.A.14)
$$+ \sum_{j=1}^{b_T}\kappa(j, b_T)\sum_{t=1+j}^{T}\Big(\boldsymbol{d}_{t+\tau}\boldsymbol{d}_{t+\tau-j}' + \boldsymbol{d}_{t+\tau-j}\boldsymbol{d}_{t+\tau}'\Big)\Bigg],$$

where $\{b_T\}$ is an integer-valued truncation point sequence satisfying $b_T \to \infty$ as $T \to \infty$ and $b_T = o(T)$ (Newey and West, 1987) in the unconditional case, and $b_T = \tau - 1$ in the conditional case. Furthermore, $\kappa(\cdot, \cdot)$ is a real-valued kernel weight function satisfying the condition

---

[37]Note that that in the unconditional case $\boldsymbol{h}_t = 1$ for all t.

that $\kappa(j, b_T) \rightarrow 1$ as $T \rightarrow \infty$ for each $j = 1, \ldots, b_T$ (Andrews, 1991), and $\kappa(j, b_T) = 0$ for $j > b_T$. We refer to Clark and McCracken (2013) for a recent review of data driven bandwidth selection methods.

Along the lines of the construction of the conditional test with $\tau = 1$, we construct the following Wald statistic which can be used in testing either unconditional or multi-step conditional equal predictive ability. The test statistic is given by

(IA.A.15)
$$S_{h,\tau} = T\overline{\boldsymbol{d}}\widetilde{\boldsymbol{\Sigma}}_T^{-1}\overline{\boldsymbol{d}},$$

where $\overline{\boldsymbol{d}} = T^{-1}\sum_{t=1}^T \boldsymbol{d}_{t+\tau}$. Before turning the properties of the proposed test statistic, we will need a slight modification of the assumptions from the previous section on one-step ahead forecasting.

Assumption 1*. *$\{\boldsymbol{h}_t\}$ and $\{\boldsymbol{w}_t\}$ are $\phi-$mixing with $\phi(t) = O\left(t^{-r/(2r-2)-\iota}\right)$, $r \geq 2$, or $\alpha-$mixing with $\alpha(t) = O\left(t^{-\frac{r}{r-2}-\iota}\right)$, $r > 2$, for some $\iota > 0$.*

Assumption 2*. *$\mathbb{E}[|\boldsymbol{d}_{t+\tau,i}|^{r+\delta}] < \infty$ for some $\delta > 0$, $i = 1, \ldots, qp$, and for all t, where subscript i indicates the ith element of $\boldsymbol{d}_{t+1}$.*

Assumption 3*. *$\boldsymbol{\Sigma}_T \equiv T^{-1}\sum_{t=1}^T \mathbb{E}[\boldsymbol{d}_{t+\tau}\boldsymbol{d}'_{t+\tau}] + T^{-1}\sum_{j=1}^{b_T}\sum_{t=1+j}^T \left(\mathbb{E}[\boldsymbol{d}_{t+\tau}\boldsymbol{d}'_{t+\tau-j}] + \mathbb{E}[\boldsymbol{d}_{t+\tau-j}\boldsymbol{d}'_{t+\tau}]\right)$ is uniformly positive definite, where $b_T = \tau - 1$ in the conditional case and $b_T = T - 1$ in the unconditional case.*

Analogues to Theorem 1, $S_{h,\tau}$ is asymptotically chi-squared distributed with qp degrees of freedom under the null hypothesis, has power under the alternative hypothesis, and is permutation invariant. We summarize these results in Theorem 2 below.

Theorem 2 (Multistep multivariate predictive ability tests). *Suppose that Assumptions 1*-3* hold.*

*A. **Asymptotic distribution under the null**. Suppose that either $\mathcal{G}_t = \{\emptyset, \Omega\}$ and $\tau \geq 1$*

or $\mathcal{F}_t \subseteq \mathcal{G}_t$ and $\tau > 1$. For any test function sequence $\{\boldsymbol{h}_t\}$, $m < \infty$, and under $\mathbb{H}_0$ in (4),

(IA.A.16)
$$S_{h,\tau} \xrightarrow{d} \chi^2(qp), \quad as \ T \to \infty.$$

B. ***Consistency under the alternative****. For any $c \in \mathbb{R}_+$ and under $\mathbb{H}_{A,h}$ in (8),*

(IA.A.17)
$$\mathbb{P}[S_{h,\tau} > c] \to 1, \quad as \ T \to \infty.$$

C. ***Permutation invariance****. Let $\boldsymbol{L}^*_{t+\tau}$ be an arbitrary permutation of the forecast losses, and define $\Delta\boldsymbol{L}^*_{t+\tau} = \boldsymbol{D}\boldsymbol{L}^*_{t+\tau}$, $\overline{\boldsymbol{d}}^* = T^{-1}\sum_{t=1}^{T}\boldsymbol{d}^*_{t+\tau}$ with $\boldsymbol{d}^*_{t+\tau} = \boldsymbol{h}_t \otimes \Delta\boldsymbol{L}^*_{t+\tau}$ and $\widetilde{\boldsymbol{\Sigma}}^*_T$ be the associated covariance estimator defined in (IA.A.14). Then,*

(IA.A.18)
$$S^*_{h,\tau} \equiv T\overline{\boldsymbol{d}}^{*'}_m(\widetilde{\boldsymbol{\Sigma}}^*_T)^{-1}\overline{\boldsymbol{d}}^* = S_{h,\tau}$$

for all T.

Consequently, a multivariate test for equal conditional multistep predictive ability or (multistep) unconditional predictive ability can be conducted by rejecting the null hypothesis whenever $S_{h,\tau} > z_{1-\alpha,qp}$, noting that the unconditional test has $q = 1$. The permutation invariance result in Theorem 2 for the unconditional case is similar to Proposition 2 in Mariano and Preve (2012), but holds under the milder Assumptions 1*–3*, and hence also applies in a setting with non-stationary data, inclusion of nested models, and explicit account of estimation uncertainty.

IA.A.2.1. Proof of Theorem 2

We start by proving part A. We proceed by a similar procedure as in the proof of Theorem 1, however with modifications due to the dependency in $\boldsymbol{d}_{t+\tau}$ under the null hypothesis. First, by Assumptions 2* and 3*, $\boldsymbol{\Sigma}_T$ is finite and uniformly positive definite. Let $\boldsymbol{\lambda} \in \mathbb{R}^{qp}$ with $\boldsymbol{\lambda}'\boldsymbol{\lambda} = 1$ and consider $\boldsymbol{\lambda}'\boldsymbol{\Sigma}_T^{-1/2}\sqrt{T}\overline{\boldsymbol{d}} = T^{-1/2}\sum_{t=1}^{T}\boldsymbol{\lambda}'\boldsymbol{\Sigma}_T^{-1/2}\boldsymbol{d}_{t+\tau}$. Furthermore, identical arguments as

in Theorem 1 imply that $\{\boldsymbol{\lambda}'\boldsymbol{\Sigma}_T^{-1/2}\boldsymbol{d}_{t+\tau}\}$ being mixing of the same size as $\{\boldsymbol{h}_t\}$ and $\{\boldsymbol{w}_t\}$. Moreover, the asymptotic variance satisfies $\sigma_d^2 = \mathrm{Var}[\boldsymbol{\lambda}'\boldsymbol{\Sigma}_T^{-1/2}\sqrt{T}\bar{\boldsymbol{d}}] = \boldsymbol{\lambda}'\boldsymbol{\Sigma}_T^{-1/2}\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_T^{-1/2}\boldsymbol{\lambda} = 1$ for all T sufficiently large. By Minkowski's inequality and computations as in (IA.A.7), $\boldsymbol{\lambda}'\boldsymbol{\Sigma}_T^{-1/2}\boldsymbol{d}_{t+\tau}$ has absolute $2+\delta$ moment for some $\delta > 0$. Then, by Corollary 3.1 in Wooldridge and White (1988) we deduce that $\boldsymbol{\lambda}'\boldsymbol{\Sigma}_T^{-1/2}\sqrt{T}\bar{\boldsymbol{d}} \xrightarrow{d} N(0,1)$. Hence, by the Cramér-Wold device it follows that $\boldsymbol{\Sigma}_T^{-1/2}\sqrt{T}\bar{\boldsymbol{d}} \xrightarrow{d} N(0, \boldsymbol{I}_{qp})$. It remains to be shown that $\tilde{\boldsymbol{\Sigma}}_T - \boldsymbol{\Sigma}_T \xrightarrow{\mathbb{P}} 0$. Consider

$$
\begin{aligned}
\tilde{\boldsymbol{\Sigma}}_T - \boldsymbol{\Sigma}_T = {} & \frac{1}{T}\sum_{t=1}^{T}\left(\boldsymbol{d}_{t+\tau}\boldsymbol{d}_{t+\tau}' - \mathbb{E}[\boldsymbol{d}_{t+\tau}\boldsymbol{d}_{t+\tau}']\right) \\
& + \frac{1}{T}\sum_{j=1}^{b_T}\kappa(j,b_T)\sum_{t=1+j}^{T}\left(\boldsymbol{d}_{t+\tau}\boldsymbol{d}_{t+\tau-j}' - \mathbb{E}[\boldsymbol{d}_{t+\tau}\boldsymbol{d}_{t+\tau-j}']\right. \\
& \left. + \boldsymbol{d}_{t+\tau-j}\boldsymbol{d}_{t+\tau}' - \mathbb{E}[\boldsymbol{d}_{t+\tau-j}\boldsymbol{d}_{t+\tau}']\right).
\end{aligned}
$$

(IA.A.19)

By Theorem 3.49 in White (2001), $\{\boldsymbol{d}_{t+\tau}\boldsymbol{d}_{t+\tau-j}'\}$ is mixing of the same size as $\{\boldsymbol{h}_t\}$ and $\{\boldsymbol{w}_t\}$ for each $j = 0, \ldots, b_T$. Moreover, each of its elements are bounded uniformly in t by Assumption 2*. Hence, since $\kappa(j,b_T) \to 1$ as $T \to \infty$ and $\kappa(0,b_T) = 1$ it follows via Corollary 3.48 in White (2001) that

$$
\frac{1}{T}\kappa(j,\tau)\sum_{t=1+j}^{T}\left(\boldsymbol{d}_{t+\tau}\boldsymbol{d}_{t+\tau-j}' - \mathbb{E}[\boldsymbol{d}_{m,t+\tau}\boldsymbol{d}_{t+\tau-j}']\right) \xrightarrow{\mathbb{P}} 0,
$$

for each $j = 0, \ldots, b_T$. Combined with (IA.A.19), this implies that $\tilde{\boldsymbol{\Sigma}}_T - \boldsymbol{\Sigma}_T \xrightarrow{\mathbb{P}} 0$ (see also Andrews (1991)). Hence, we can deduce via similar steps as in (IA.A.9) that $S_{h,\tau} \xrightarrow{d} \chi^2(qp)$ as $T \to \infty$. $\qquad\square$

We now prove part B. The result follows by arguments similar to those in the proof of Theorem 1. Hence, $\{\boldsymbol{d}_{t+\tau}\}$ is mixing with the same size as $\{\boldsymbol{h}_t\}$ and $\{\boldsymbol{w}_t\}$ and each element in $\boldsymbol{d}_{t+\tau}$ is bounded uniformly in t by Assumption 2*. Then it follows by Corollary 3.48 in White (2001) that $\bar{\boldsymbol{d}} - \mathbb{E}[\bar{\boldsymbol{d}}] \xrightarrow{\mathbb{P}} 0$, and consequently similar computations as in (IA.A.11)

applies. By arguments identical to those in the proof of part A., $\tilde{\boldsymbol{\Sigma}}_T - \boldsymbol{\Sigma}_T \overset{\mathbb{P}}{\to} 0$, where $\boldsymbol{\Sigma}_T$ is positive definite by Assumption 3*. Theorem 8.13 in White (1994) then implies that under $\mathbb{H}_{A,h}$ in (8) and for any constant $c \in \mathbb{R}_+$, $\mathbb{P}[S_{h,\tau} > c] \to 1$ as $T \to \infty$. □

Last, we prove part C. Due the arguments in the proof of Theorem 1, it suffices to show that $\tilde{\boldsymbol{\Sigma}}_{T*} = \boldsymbol{A}\tilde{\boldsymbol{\Sigma}}_T\boldsymbol{A}'$, where $\boldsymbol{A} = \boldsymbol{I}_q \otimes \boldsymbol{B}$. Thus, let

$$\tilde{\boldsymbol{\Sigma}}_T(b) \equiv \frac{1}{T}\sum_{t=1+b}^{T} \boldsymbol{d}_{t+\tau}\boldsymbol{d}'_{t+\tau-b},$$

for $b = 0, 1, 2 \ldots$. It then follows that

$$\tilde{\boldsymbol{\Sigma}}_T(b)^* \equiv \frac{1}{T}\sum_{t=1+b}^{T} \boldsymbol{d}^*_{t+\tau}\boldsymbol{d}^{*'}_{t+\tau-b} = \frac{1}{T}\sum_{t=1+b}^{T} \boldsymbol{A}\boldsymbol{d}_{t+\tau}\boldsymbol{d}'_{t+\tau-b}\boldsymbol{A}' = \boldsymbol{A}\tilde{\boldsymbol{\Sigma}}_T(b)\boldsymbol{A}'.$$

Consequently, it follows that $\tilde{\boldsymbol{\Sigma}}_T^* = \boldsymbol{A}\tilde{\boldsymbol{\Sigma}}_T\boldsymbol{A}'$, which completes the proof. □

IA.B. A check of size and power properties

This section performs a Monte Carlo study to investigate the finite sample properties of our multivariate test for equal conditional predictive ability. We focus on size and power properties of the test in a setting that closely resembles the empirical analyses in both a full out-of-sample analysis and when used in the dynamic ranking rule.[38]

We examine a situation where the forecasts have equal predictive ability unconditionally, but conditional on some state variable $\widetilde{h}_t$ at least one of the forecasts are more (or less) accurate than the others. The data-generating process is set to

$$(IA.B.20) \qquad \Delta \boldsymbol{L}_{t+1} = \boldsymbol{\mu} \left(\widetilde{h}_t - \varrho \right) + \boldsymbol{\varepsilon}_{t+1},$$

where $\mathbb{P}\left[\widetilde{h}_t = 1\right] = \varrho$ and $\mathbb{P}\left[\widetilde{h}_t = 0\right] = 1 - \varrho$. To allow for the presence of estimation error (approximately) asymptotically, as delineated by our theoretical setting, we re-sample with replacement from de-meaned loss differentials from our empirical analysis when generating $\boldsymbol{\varepsilon}_{t+1}$. In this way, they retain the influence of the estimation coming the forecasting models as well as ensuring simulated time series that exhibit realistic empirical behavior. Note also that $\mathbb{E}[\Delta \boldsymbol{L}_{t+1}] = \boldsymbol{0}$, together with $\mathbb{E}[\Delta \boldsymbol{L}_{t+1}|\widetilde{h}_t = 1] = \boldsymbol{\mu}(1 - \varrho)$ and $\mathbb{E}[\Delta \boldsymbol{L}_{t+1}|\widetilde{h}_t = 0] = -\boldsymbol{\mu}\varrho$. That is, the unconditional null hypothesis is true, whilst the conditional one is not necessarily so, depending on the value of (the elements in) $\boldsymbol{\mu}$ and ϱ.

We consider three sample size lengths: short, medium, and long. The medium size equals the length of our full out-of-sample window, $T = 348$, the short size equals the sample length used in the dynamic ranking rule in the application, $T_1 = 120$, and the long size is set to 1,000 observations. Consistent with our empirical analysis, we set $p = 5$ as the number of models under comparison less one due to the computation of loss differentials. Since our

---

[38]We refer to Borup and Thyrsgaard (2017) for extensive Monte Carlo evidence for all test statistics with remedies for large dimensions and power enhancement techniques.

ranking rules eliminate a model sequentially until it no longer rejects, we consider the full range $p = 1, \ldots, 5$. When $p < 5$, we randomly sample without replacement (in any random order) among our full set of models and subsequently reconstruct loss differentials based on the selected models. Note that any reshuffling of the order of models has no influence on the test statistic due to its permutation invariance, presented in Theorem 1, such that it has no influence on the performance of the test statistic within a fixed p. To obtain (samples of) $\boldsymbol{\varepsilon}_{t+1}$, we consider the empirical loss differentials coming from forecasting the 2-year and the 5-year bond, respectively, as illustrative cases. We set $\varrho = 0.4$, since this links to our findings below that documents notable superior predictability of at least one model in each of the high and low economic activity or uncertainty states, and less differences in predictive accuracy within the normal state. We use 10,000 Monte Carlo replications.

IA.B.1.   Size properties

To examine the size properties of our test, we set $\boldsymbol{\mu} = \boldsymbol{0}$ such that both the unconditional and conditional null hypothesis are true. We consider two implementations of the test. The first is unconditional and uses $\boldsymbol{h}_t = 1$ for all t, whereas a conditional implementation uses $\boldsymbol{h}_t = (1, \tilde{h}_t)'$. The results are reported in Table IA.1 for a significance level of 5%. Conclusions are identical using a 1% and 10% significance level, and relevant tables are available upon request.

[Insert Figure IA.1 About Here]

It is clear that both the conditional and unconditional tests are well-sized, showing negligible deviations from the nominal significance level. Those minor deviations generally decrease in sample size and increase in number of models under comparison. It is comforting to note that the tests maintain good size properties for the short sample size used in the dynamic ranking rule. There is no notable difference when sampling from loss differentials associated with the 2-year or 5-year bonds, except from in the short sample case where the 5-year bond loss differentials lead to a slight undersizing.

IA.B.2.   Power properties

To examine the power properties of our test, we let the first element of $\boldsymbol{\mu}$ deviate from zero, and set the remaining elements equal to zero, similarly to Mariano and Preve (2012). Denote this first element by μ_1. The deviation is anchored in the empirical loss differentials, making it realistic in the context of the present paper. Specifically, we compute the average absolute loss differentials across all models within the low and high economic activity states using PMI defined in the empirical section of the main paper, denoting it by $\hat{\eta}$. We then set $\mu_1 = c\hat{\eta}$ where $c \in [0, 2.5]$.[39] Given the specification in (IA.B.20) and $\varrho = 0.4$, this allows μ_1 to deviate at most 1.5 times the empirical value of average absolute loss differentials. We have also implemented a version that lets all elements of $\boldsymbol{\mu}$ deviate from zero with a fraction c of each respective element's average absolute loss differentials within the low and high activity states. The power is uniformly stronger in this case, and results are available upon request. Note also that, in both versions, the unconditional null hypothesis remains true. We therefore set $\boldsymbol{h}_t = (1, \widetilde{h}_t)'$ and examine the power of the conditional version of our equal predictability test. The power curves for a 5% significance level are depicted in Figure IA.1. Conclusions are identical using a 1% and 10% significance level, and the results are available upon request.

[Insert Figure IA.1 About Here]

In line with the theoretical power result in Theorem 1, the test is consistent under the (local) alternative considered, as power increases to unity for stronger deviations from the null. It correctly exhibits empirical rejections equal to the nominal size at $c = 0$. Power is stronger for fewer model comparisons, as expected, but the difference is not substantial. Similar to size properties, it is comforting that the test exhibits good power properties even for the relatively short sample length. To put this into context, for $c = 1/\varrho = 1.67$, we recover the empirical value of the mean absolute values of loss differentials obtained in the empirical

---

[39]We also ran the simulation using $\mathcal{U}$ as state variable, yielding similar conclusions, yet somewhat stronger power.

analysis when using (IA.B.20). In this case, the power exceeds 0.94 for the smallest sample size and $p = 5$, showing very desirable power properties. There is no notable difference when sampling from loss differentials associated with the 2-year or 5-year bonds.

IA.C. Bond data

IA.C.1. Bond risk premia

We use the Gürkaynak et al. (2007) dataset from 1962:M1 to 2018:M12. The time t log yield on a k-period bond is computed using the methods developed in Nelson and Siegel (1987) and Svensson (1994) as

$$y_t^{(k)} = \beta_{0,t} + \beta_{1,t} \frac{1 - \exp\left(-\frac{n}{\kappa_{1,t}}\right)}{\frac{n}{\kappa_{1,t}}} + \beta_{2,t} \left[\frac{1 - \exp\left(-\frac{n}{\kappa_{1,t}}\right)}{\frac{n}{\kappa_{1,t}}} - \exp\left(-\frac{n}{\kappa_{1,t}}\right) \right]$$

$$\text{(IA.C.21)} \qquad + \beta_{3,t} \left[\frac{1 - \exp\left(-\frac{n}{\kappa_{2,t}}\right)}{\frac{n}{\kappa_{2,t}}} - \exp\left(-\frac{n}{\kappa_{2,t}}\right) \right],$$

where we use parentheses in the superscript to distinguish maturity from exponentiation and $n = \frac{k}{m}$ and m denotes, respectively, the bond maturity in years and the number of periods per year.

Let $p_t^{(k)} = -\left(\frac{k}{m}\right) y_t^{(k)}$ be the log price of a k-period bond at time t. The log forward rate at time t for loans between $t + k - 1$ and $t + k$ is defined as

$$\text{(IA.C.22)} \qquad f_t^{(k)} = p_t^{(k-1)} - p_t^{(k)} = -\frac{k-1}{m} y_t^{(k-1)} + \frac{k}{m} y_t^{(k)}.$$

The excess return to purchasing a k-period bond today and selling it as a $k - 1$ period bond after one month is

$$\text{(IA.C.23)} \qquad rx_{t+1}^{(k)} = p_{t+1}^{(k-1)} - p_t^{(k)} - p_t^{(1)} = -\frac{k-1}{m} y_{t+1}^{(k-1)} + \frac{k}{m} y_t^{(k)} - \frac{1}{m} y_t^{(1)},$$

where $y_t^{(1)}$ denotes the risk-free one-period rate that we proxy using the implied yield on a one-month Treasury bill obtained from the Center for Research in Security Prices (CRSP) as

in Gargano et al. (2019).[40]

IA.C.2.   Predictor variables

We consider a set of standard bond predictors from the extant literature. In particular, we consider yield spreads (Campbell and Shiller, 1991), forward spreads (Fama and Bliss, 1987), principal components of yields (Litterman and Scheinkman, 1991), forward rates (Cochrane and Piazzesi, 2005), the output gap (Cooper and Priestley, 2009), macroeconomic factors (Ludvigson and Ng, 2009), and interest rate cycles (Cieslak and Povala, 2015). We discuss the construction of each of the variables below.

The Campbell-Shiller (CS) yield spreads are computed for each bond maturity as

$$(IA.C.24) \qquad\qquad ys_t^{(k)} = y_t^{(k)} - y_t^{(1)},$$

where $y_t^{(k)}$ denotes the time t log yield on a bond with k periods to maturity and $y_t^{(1)}$ denotes the safe one-period return measured using the implied yield on a one-month Treasury bill obtained from CRSP. The Fama-Bliss (FB) forward spreads are computed similarly as

$$(IA.C.25) \qquad\qquad fs_t^{(k)} = f_t^{(k)} - y_t^{(1)},$$

where $f_t^{(k)}$ denotes the forward rate for loans between $t + k - 1$ and $t + k$. The principal component (PC) of yields are computed from bonds with 12, 24, 36, 48, and 60 months left to maturity and we focus on the first three components often referred to as level, slope, and curvature. These components account for almost all of the variation in yields. The Cochrane-Piazzesi (CP) single factor is formed from a linear combination of forward rates

---

[40]For $k = 1$, we have that $f_t^{(1)} = y_t^{(1)}$ and that $y_t^{(k-1)} = y_t^{(0)} = 0$ due to $p_t^{(0)}$ being zero (log of terminal payoff of one is zero).

using the projection

$$(IA.C.26) \qquad \overline{rx}_{t+1} = \delta + \gamma_1 f_t^{(12)} + \gamma_2 f_t^{(24)} + \gamma_3 f_t^{(36)} + \gamma_4 f_t^{(48)} + \gamma_5 f_t^{(60)} + \varepsilon_{t+1},$$

where $\overline{rx}_{t+1} = \frac{1}{4} \sum_{i=2}^{5} rx_{t+1}^{(i \times 12)}$ can be viewed as the excess return on a portfolio of Treasury bonds with different maturities. The CP factor is obtained as the fitted values from the regression, $\text{CP}_t = \widehat{\delta} + \widehat{\boldsymbol{\gamma}} \boldsymbol{f}_t$, with $\widehat{\boldsymbol{\gamma}} = (\widehat{\gamma}_1, \widehat{\gamma}_2, \widehat{\gamma}_3, \widehat{\gamma}_4, \widehat{\gamma}_5)$ and $\boldsymbol{f}_t = (f_t^{(12)}, f_t^{(24)}, f_t^{(36)}, f_t^{(48)}, f_t^{(60)})'$. The Ludvigson-Ng (LN) macroeconomic factor is based on a large $T \times M$ panel of macroeconomic variables, x, that we assume can be adequately described by a static factor model

$$(IA.C.27) \qquad x_{i,t} = \kappa_i g_t + \nu_{i,t,,}$$

where g_t is an $s \times 1$ vector of common factors with $s \ll M$ that we estimate using principal component analysis. We use the comprehensive FRED-MD dataset from McCracken and Ng (2016). Following Ludvigson and Ng (2009), we build a single factor as a linear combination of a subset of the principal components. We determine the subset using the BIC and obtain the factor from a projection of $\overline{rx}_{t+1}$ onto the set of selected macroeconomic factors.

[Insert Table IA.2 About Here]

Table IA.2 presents descriptive statistics for the set of predictors (Panel A) along with contemporaneous correlations (Panel B). All variables are constructed using the full range of available observations here, but are constructed recursively in the out-of-sample exercise. Yield spreads and forward spreads are fairly persistent with first-order autocorrelations between 0.82 and 0.92 and are heavily cross-correlated. Unsurprisingly, PC2 — the slope component of the yield curve — is strongly related to both yield and forward spreads. CP and LN are similarly positively correlated with the spread variables and also positively correlated with each other. Last, we note that CP and LN are relatively less persistent compared to the remaining variables.

IA.C.3. Additional details on state variables

This section provides additional details on the state variables used in the empirical analyses to address state-dependencies in bond return predictability.

The Purchasing Managers' Index (PMI) is a leading economic indicator constructed from monthly surveys of the manufacturing sector and is released on the first business day of every month. The survey covers senior purchasing executives (or similar) at around 400 manufacturing companies. The PMI is a weighted average of the following five indices: New Orders (30%), Output (25%), Employment (20%), Suppliers' Delivery Times (15%) and Stocks of Purchases (10%). For the PMI calculation, the Suppliers' Delivery Times Index is inverted so that it moves in a comparable direction to the other indices. For more details, we refer to IHS Markit's product page (https://ihsmarkit.com/products/pmi.html). The index ranges from 0 to 100 and is specifically designed to capture the state of the economy with values above (below) 50 indicating an overall increase (decrease) compared to the previous month. We have collected the index from Global Financial Data.

The macroconomic uncertainty index ($\mathcal{U}$) of Jurado et al. (2015) measures a common component in the time-varying volatilities of h-step ahead forecast errors across a large number of macroeconomic series that include categories such as real activity, prices, and financial assets. The idea build on the premise that what matters for economic decision making is not whether particular economic indicators have become more or less variable or disperse *per se*, but rather whether the economy has become more or less predictable. The index is therefore associated with the variance of the unpredictable components of macroeconomic variables. The index is obtained from Sydney Ludvigson's website (https://www.sydneyludvigson.com/macro-and-financial-uncertainty-indexes).

[Insert Table IA.3 About Here]

Table IA.3 presents full sample descriptive statistics for our two state variables that captures economic activity and uncertainty, respectively: the Purchasing Managers' Index

(PMI) and the macroeconomic uncertainty index of Jurado et al. (2015). The series are both highly persistent with autocorrelation coefficients well above 0.9. Most importantly, we note that the series obtains a negative contemporaneous correlation of -0.48 in the data, suggesting that they capture part of the some features, but are not perfect substitutes.

IA.D. Additional empirical results

IA.D.1. In-sample predictive regressions

Table IA.4 presents full sample least squares estimation results to facilitate comparison with the extant literature. Specifically, we estimate predictive regressions of the form presented in (1) with the risk premium on a Treasury bond with k-periods to maturity $rx_{t+1}^{(k)}$ as the dependent variable. We focus on bonds with 24, 36, 48, and 60 months to maturity and consider models based on the predictor variables outlined in Section B. We stress that these results are not available to a real-time investor, but they are useful for gauging the mechanisms of the predictive models.

[Insert Table IA.4 About Here]

The slope coefficients for CS and FB are all positive and increasing with maturity and are all statistically significant at conventional levels.[41] We note that these positive slope coefficients imply negative slopes for the companion regression of yield or forward spreads on future yield changes as documented in Campbell and Shiller (1991). Thus, both yield and forward spreads contain useful information about future bond excess returns over the full range of available observations. Turning to the principal components, we find that PC1 has a constant slope coefficient across maturities, PC2 increases monotonically, and PC3 displays an inverse U-shape. PC1 and PC3 are mostly insignificant, whereas PC2 is significant for the longer maturities. This mirrors the results for CS, but shows that maturity-specific spreads are more informative than the common slope factor. Last, CP and LN both display monotonically increasing slope coefficients that are highly significant. Of all the models, LN

---

[41]Bauer and Hamilton (2018) show that statistical test of predictive regression in full sample analyses are subject to serious small sample distortions when using 12-month overlapping returns. However, we use one-month non-overlapping returns here and are therefore not affected by their results. See also the discussion in Gargano et al. (2019).

xxi

appears to explain the largest fraction of bond risk premia, closely followed by CP and yield spreads. Overall, in-sample results points to predictive relation between all our candidate predictors.

IA.D.2.   Decision rule and model selection

Figure IA.2 illustrates the models selected for the best set of models using the decision rule over time using PMI and $\mathcal{U}$ as conditioning variables, respectively. Green (yellow) shaded aras indicate high (low) states identified using the 20% and 80% quantiles of the series. A "+" indicates inclusion.

[Insert Figure IA.2 About Here]

[Insert Figure IA.3 About Here]

Figures IA.3 illustrates the size of the set of best models selected over time using the decision rule using PMI and $\mathcal{U}$ as conditioning variables, respectively. We note that the best set of models varies considerably over time and includes situations in which the set include all models, leading to forecasts equal to EW, and situations with a singleton. That is, at times there is no need for trimming of the full set of models and at other times we should only use the forecasts from a single model. Importantly, this tells us why dynamically trimming leads to improvements over a simple, static forecast combination rule.

IA.D.3.   Full out-of-sample period ranking

We here device a rule that uses the full out-of-sample period for ranking all forecasting methods based on their predictive accuracy conditional on the state variable. We formulate the rule using a single state variable (and a constant), but note that the rule can be extended directly to a setting with multiple state variables. Since $\widetilde{h}_t$ may be continuous, we assume that it can be classified into a finite set of $\mathcal{A}$ discrete, non-empty, states s_a, $a = 1, \ldots, \mathcal{A}$. For

example, the state variable can be a measure of economic growth, which may be classified into recessionary or expansionary states, or a measure of macroeconomic uncertainty, which may be classified into low, medium, and high uncertainty states.

Let M^0 denote the set of $p+1$ forecasting methods under consideration and M_a^* a set of best forecasting methods in terms of some loss function within the ath state. We then consider the following three-step procedure.

Step 0: Set $M_a = M^0$ for $a = 1, \ldots, \mathcal{A}$. For all pairwise combinations of of forecasting methods, $j, i \in M_t, i \neq j$, estimate by OLS the regression model

$$(\text{IA.D.28}) \qquad \Delta L_{t+1}^{i,j} = L_{t+1}^i - L_{t+1}^j = \boldsymbol{\varphi}^j \boldsymbol{h}_t + \eta_{t+1}$$

over the entire out-of-sample period. The conditional expectation of the loss differentials within each state, $\mathbb{E}\left[\Delta L_{t+\tau}^j | s = s_a\right] = \varphi_0^j + \varphi_1^j \mathbb{E}\left[\widetilde{h_t} | s = s_a\right]$, $a = 1, \ldots, \mathcal{A}$, is approximated by $\widehat{\varphi}_0^j + \widehat{\varphi}_1^j \widehat{\mu}_{\widetilde{h}}^a$, where $\widehat{\mu}_{\widetilde{h}}^a$ is the sample average of the state variable $\widetilde{h}_t$ in state s_a. Based on the estimated conditional means, rank all $p+1$ methods (using a normalization of one method) in all states. The forecasting method with lowest predicted loss across all pairwise combinations is ranked first and the method with highest predicted loss is ranked last.

Step 1: Run the multivariate test for equal conditional predictive ability.

Step 2: If the test is not rejected, set $M_a^* = M_a$. Otherwise, eliminate the lowest ranked forecasting method from M_a based on the ranking that associates with state a. Iterate Steps 1–2 until the null is no longer rejected for all $\mathcal{A}$ states.

Concluding the algorithm leads to a set M_a^* for each state s_a that contains the best forecasting methods statistically indistinguishable in terms of predictive ability.

To facilitate interpretation and consistent with the empirical analysis in the main paper, we classify the continuous state variables, PMI and $\mathcal{U}$, into low, normal, and high economic

activity (uncertainty) states using the 20% and 80% quantiles as discussed in Section C.

[Insert Figure IA.4 About Here]

Figure IA.4 illustrates the full out-of-sample elimination order of the predictive models determined by the ranking rule discussed in Section III when conditioning on low, normal, and high PMI and $\mathcal{U}$ states, respectively, using a 10% significance level. We uncover a striking pattern over the full sample. In particular, the EH tends to be eliminated in high (low) economic activity (uncertainty) states across the entire maturity spectrum. Interpreting the EH as a no-predictability benchmark implies that bond risk premia are predictable when economic activity (uncertainty) is high (low). Conversely, the consistent inclusion of EH in low (high) economic activity (uncertainty) states suggests a lack of predictability. Overall, we argue that our empirical results are consistent with, and clearly points to, state-dependencies in bond excess return predictability linked to economic activity and uncertainty. Bond excess returns are predictable in states with high (low) economic activity (uncertainty), whereas the EH serves as a reliable anchor in the remaining states of the world.

IA.D.4.   Links to uncertainty

Table IA.6 presents contemporaneous correlations among $\mathcal{U}$ and the risk premia estimates from the set of individual models, EW, and the dynamic forecast combinations generated by PMI, $\mathcal{U}$, and NONE.

[Insert Table IA.6 About Here]

We find that most forecasts are positively correlated with uncertainty, implying that investors higher risk premia in periods with heightened uncertainty. The exception is CS and FB for the shorter maturities, where we observe negative correlations. As for our main results concerning the relation to economic activity (see Table 6), we find that LN displays the highest correlation with $\mathcal{U}$ among the individual predictors and EW. Turning to the dynamic forecast combination estimates in Panel B, we find that both PMI and $\mathcal{U}$ trimming delivers forecasts

that are tightly linked to uncertainty. That is, not only do they produce countercyclical risk premia estimates, they only procedure forecasts closely linked to uncertainty.

IA.D.5.   Additional results for economic value

Figure IA.5 plots the cumulative CER gains for the individual predictor variables along with the equal-weighted forecast (EW). The results largely mirrors those in Table 7 in the main paper and illustrate that most individual predictors fail to deliver economic value on a consistent basis. The exception being LN.

[Insert Figure IA.5 About Here]

[Insert Table IA.7 About Here]

Table IA.7 reconstructs the results from Table 7 in the main paper using instead a coefficient of relative risk aversion of $\gamma = 5$ to verify that our results are robust to other, and lower, values of risk aversion. The table clearly demonstrates that this is the case.

As another test of the improvement in economic value we assess how the predictive content in our models under consideration maps into Sharpe ratio improvements. We follow the approach by Gu et al. (2020) and use that the Sharpe ratio $(SR_{i,k}^*)$ earned by an active investor that utilizes predictive information (summarized by the $R_{OS,i,k}^2$) for model i and bond maturity k is given by

$$SR_{i,k}^* = \sqrt{\frac{SR_k^2 + R_{OS,i,k}^2}{1 - R_{OS,i,k}^2}},$$

where SR_k is the Sharpe ratio earned by a buy-and-hold investor on a k-maturity bond. We then report annualized Sharpe ratio improvements for each method, $\sqrt{12}\left(SR_{i,k}^* - SR_k\right)$, in the cases where $R_{OS,i,k}^2 \geq 0$. These are collected in Table IA.8.

[Insert Table IA.8 About Here]

It is clear that individual predictors almost always lead to Sharpe ratio reductions and if any gain is achieved it is generally small. On the other hand, EW provides notable Sharpe ratio improvements, yet our dynamic forecast combination is superior for all bond maturities. For instance, the buy-and-hold Sharpe ratio of a 2-year bond, which is 0.90 in our out-of-sample period, can be improved with 0.48 (0.59) using PMI ($\mathcal{U}$) as state variable. Improvements for NONE are almost identical to EW.

IA.E. Description of the non-linear model and estimation

IA.E.1. The model

We consider a one-factor version of the model in Feldhütter et al. (2018). In the model, uncertainty is driven by an one-dimensional Brownian motion $W(t)$. There is then an one-factor gaussian factor $X(t)$ with dynamics

$$(\text{IA.E.29}) \qquad dX(t) = \kappa(\theta - X(t))dt + \Sigma dW(t),$$

where θ, κ and Σ are constants. In the absence of arbitrage, there exist a stochastic discount factor (SDF) given by

$$(\text{IA.E.30}) \qquad M(t) = M_0(t)\left(1 + \gamma e^{-\beta X(t)}\right),$$

where γ is a non-negative constant, β is a constant, and $M_0(t)$ is a strictly positive process. This implies that $M(t)$ is strictly positive. Note that (IA.E.30) is a departure from standard term structure models in that this model specify the functional form of the SDF directly rather than pinning it down from short rate and market price of risk specifications.

The price $P(t,T)$ of a zero-coupon bond at time t that matures at T is given by

(IA.E.31)
$$P(t,T) = s(t)P_0(t,T) + (1 - s(t))P_1(t,T),$$

where

(IA.E.32)
$$s(t) = \frac{1}{1 + \gamma e^{-\beta X(t)}},$$

(IA.E.33)
$$P_i(t,T) = e^{A_i(T-t) + B_i(T-t)X(t)}.$$

The coefficients $A_i(T-t)$ and $B_i(T-t)$ solve the set of ordinary differential equations

(IA.E.34)
$$\frac{dB_i(T-t)}{d(T-t)} = -(\kappa + \Sigma\lambda_{i,X})B_i(T-t) - \rho_{i,X},$$

(IA.E.35)
$$\frac{dA_i(T-t)}{d(T-t)} = \frac{1}{2}\Sigma^2 B_i(T-t)^2 + B_i(T-t)(\kappa\theta - \Sigma\lambda_{i,0}) - \rho_{i,0},$$

(IA.E.36)
$$\lambda_{1,0} = \lambda_{0,0} + \Sigma\beta,$$

(IA.E.37)
$$\lambda_{1,X} = \lambda_{0,X},$$

(IA.E.38)
$$\rho_{1,0} = \rho_{0,0} + \beta\kappa\theta - \frac{1}{2}\beta^2\Sigma^2,$$

(IA.E.39)
$$\rho_{1,X} = \rho_{0,X} - \kappa\beta - \lambda_{0,X}\Sigma\beta.$$

For identification, we follow Feldhütter et al. (2018) and set $\theta = 0$ and $\Sigma = 1$. If $\beta = 0$ or $\gamma = 0$, then the model collapses to a standard Gaussian term structure model. We consider essentially affine market prices of risk.

IA.E.2. Estimation procedure

We estimate the one-factor version of the non-linear term structure model by maximum likelihood using Kalman filtering, shifting notation for expositional reasons to using subscript for time indicators. We cast the model into a state space form with a transition equation that

describes the dynamics of the state factor, X_t, and a measurement equation that describes the relationship between X_t and both yields and excess returns.

Let Y_t denote a vector of yields and excess returns, then the measurement equation is specified as

$$(\text{IA.E.40}) \qquad Y_t = z\left(\Theta, X_t\right) + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}\left(0, \Sigma_{\text{measurement}}\right)$$

where $z\left(\cdot\right)$ is the pricing function, Θ is a vector of model parameters, and ε_t is vector of i.i.d. Gaussian pricing errors with covariance matrix $\Sigma_{\text{measurement}}$. To reduce the number of parameters, we assume that pricing errors for all yields have the same variance σ^2_{yields} and that all pricing errors for excess returns have the same variance σ^2_{excess}.

The transition equation for the dynamics of the state variable X_t have the form

$$(\text{IA.E.41}) \qquad X_{t+\tau} = C + D X_t + \eta_{t+\tau}, \quad \eta_{t+\tau} \sim \mathcal{N}\left(0, Q\right)$$

where $D = e^{-\kappa\tau}$, $C = (1 - e^{-\kappa\tau})\theta$, and Q denotes the conditional variance of $X_{t+\tau}$ given X_t. The latter is constant provided that X_t is Gaussian. We refer to Fackler (2000) for a closed-form solution. Similarly, the expected variance is given as

$$(\text{IA.E.42}) \qquad \Sigma_{x,t+\tau|t} = D' \Sigma_{x,t|t} D + Q.$$

In the non-linear model, the pricing function $z\left(\cdot\right)$ is nonlinear both for yields and excess returns. The unscented Kalman filter seems to be the standard approach for gauging such non-linearities (see, among others, Cieslak and Povala (2016), Filipović, Larsson, and Trolle (2017) and Feldhütter et al. (2018)). We follow Christoffersen, Dorion, Jacobs, and Karoui (2014) and implement the square-root unscented Kalman filter of Van Der Merwe and Wan (2001) that is numerically more stable than the standard unscented Kalman filter.

The unscented Kalman filter evaluates the measurement equation in a set of sigma points

rather than linearizing it as in the standard Kalman filter. These sigma point vectors are given by

(IA.E.43) $\chi_0 = X_{t+\tau|t},$

(IA.E.44) $\chi_i = X_{t+\tau|t} + \left(\sqrt{(N+\lambda)}\sqrt{\Sigma_{x,t+\tau|t}}\right)_i \quad$ for $i = 1,\ldots,N,$

(IA.E.45) $\chi_i = X_{t+\tau|t} - \left(\sqrt{(N+\lambda)}\sqrt{\Sigma_{x,t+\tau|t}}\right)_{i-N} \quad$ for $i = N+1,\ldots,2N,$

where N is the number of state factors in the state-space system. The number of sigma points is, therefore, equal to $2N+1$. Last, λ is a scaling factor determined by

(IA.E.46) $$\lambda = \alpha^2 (N - \kappa) - N.$$

We follow Van Der Merwe and Wan (2001) and set $\kappa=0$ and $\alpha = 10^{-3}$, where α is intended to minimize higher-order effects. The first moment of the measurement equation is approximated by

(IA.E.47) $$\overline{Y}_{t+\tau|t} \approx \sum_{i=0}^{2N} W_i^\mu \mathcal{M}_i$$

where $\mathcal{M}_i = z(\Theta, \chi_i)$. The variance of the measurement equation is

(IA.E.48) $S_y = \text{cholupdate}\left(qr\left[\sqrt{W_1^\sigma}(\mathcal{M}_{1:2N} - \mathcal{M}_0)\,\text{chol}(H)\right], \mathcal{M}_0 - \overline{Y}_{t+\tau|t}, W_1^\sigma \right),$

(IA.E.49) $F_{t+\tau|t} = S_y S_y',$

where $qr(\cdot)$ is the orthogonal-triangular decomposition such that $A = qr$, where q is a orthogonal matrix and r is a upper triangular matrix. The function returns the r matrix. The function "chol" is the cholesky decomposition such that if $B = \text{chol}(A)$ then $BB' = A$. The function "cholupdate(A, B, C)" updates the cholesky decomposition A such that cholupdate$(A, B, C) = \text{chol}(AA' + CBB')$. In the case where C is a matrix, then the

procedure is performed column by column.

The weights W_i^μ and W_i^σ are given by

(IA.E.50)
$$W_0^\mu = \frac{\lambda}{N + \lambda},$$

(IA.E.51)
$$W_0^\sigma = \frac{\lambda}{N + \lambda} + 1 - \alpha^2 + \beta,$$

(IA.E.52)
$$W_i^\mu = W_i^\sigma = \frac{\lambda}{2(N + \lambda)}, \text{ for } i = 1, \dots, 2N.$$

When X_t is Gaussian, setting $\beta = 2$ is optimal (Van Der Merwe and Wan, 2001). The next step is to use these to predict the measurement equation using the updating step. The Kalman gain is

(IA.E.53)
$$K_{t+\tau} = \left(P_{xy}/S_y' \right) / S_y.$$

The update step is then performed by

(IA.E.54)
$$X_{t+\tau|t+\tau} = X_{t|t-1} + K_{t+\tau} \left(Y_{t+\tau} - \overline{Y}_{t+\tau|t} \right)$$

(IA.E.55)
$$\Sigma_{x,t+\tau|t+\tau} = \Sigma_{x,t+\tau|t} - \Sigma_{xy} F_{t+\tau|t}^{-1} \Sigma_{xy}.$$

To initialize the Kalman filter, we set $X_{0|0}$ and $\Sigma_{x,0|0}$ equal to, respectively, the unconditional mean and variance of X_t. We can then construct the loglikelihood function as

(IA.E.56)
$$l(\Theta) = \sum_{t=1}^{T} -\frac{N}{2} log(2\pi) - \frac{log\left(| F_{t|t-1} | \right)}{2} - 0.5 \left(Y(t) - \overline{Y}_{t|t-1} \right)' F_{t|t-1}^{-1} \left(Y(t) - \overline{Y}_{t|t-1} \right)$$

The estimates $\widehat{\Theta}$ can then be obtained by maximizing (IA.E.56) with respect to the parameters, Θ. We follow Cieslak and Povala (2016) and maximize the loglikelihood function using the differential evolution algorithm. The optimization is repeated with multiple plausible initial starting values.

Table IA.1: Empirical size properties

This table reports the rejection frequency (empirical size) of the multivariate test for equal predictive ability with a nominal size of 5%, data-generating process given by (IA.B.20) with $\boldsymbol{\mu} = \mathbf{0}$, and 10,000 Monte Carlo replications. We implement an unconditional test that sets $\boldsymbol{h}_t = 1$ for all t and a conditional test that sets $\boldsymbol{h}_t = (1, \tilde{h}_t)'$, and use three samples sizes referred to as short (120 observations), medium (348 observations) and long (1,000 observations). Panel A (B) reports results where $\boldsymbol{\varepsilon}_{t+1}$ in (IA.B.20) is sampled from the empirical loss differentials when forecasting the 2-year (5-year) bond. The value of p indicates the dimension of the test arising from the number of comparing models less one.

	Unconditional			Conditional		
	Short	Medium	Long	Short	Medium	Long
			Panel A: 2-year bond			
p=1	5.22	5.08	4.84	5.48	4.82	5.08
p=2	4.65	4.99	5.09	4.85	5.27	4.98
p=3	4.82	5.34	4.88	5.33	5.38	5.17
p=4	4.91	5.12	5.03	4.84	5.27	5.14
p=5	5.22	4.61	4.76	4.62	5.10	5.29
			Panel B: 5-year bond			
p=1	4.73	4.87	5.26	3.96	4.53	4.99
p=2	4.36	4.56	4.99	4.07	4.48	4.92
p=3	4.05	4.47	4.99	3.79	4.56	4.78
p=4	4.38	4.24	4.96	3.69	4.34	5.11
p=5	4.30	4.59	5.02	3.22	4.50	4.89

Table IA.2: Predictor variables

This table reports descriptive statistics (Panel A) for the predictor variables used in our empirical study and their contemporaneous correlations (Panel B). We report means, standard deviations, skewness, kurtosis, and first-order autocorrelation (AC(1)) of each predictor. The predictors include maturity-specific yield (Campbell and Shiller, 1991) and forward (Fama, 1984) spreads, the first three principal components of the yield curve (Litterman and Scheinkman, 1991), the Cochrane and Piazzesi (2005) (CP) factor, and the Ludvigson et al. (2019) (LN) factor. PC1–3, CP, and LN are based on full sample estimates. The sample period is January 1962 to December 2018.

	Campbell-Shiller				Fama-Bliss								
	2-year	3-year	4-year	5-year	2-year	3-year	4-year	5-year	PC1	PC2	PC3	CP	LN
Panel A: Descriptive statistics													
Mean	0.07	0.09	0.10	0.11	0.11	0.13	0.15	0.16	12.20	0.95	-0.17	0.14	0.14
Std. dev.	0.07	0.08	0.08	0.09	0.09	0.11	0.12	0.13	7.03	0.59	0.10	0.20	0.32
Skewness	0.88	0.41	0.13	-0.03	0.06	-0.19	-0.23	-0.17	0.48	0.11	-0.79	0.71	1.52
Kurtosis	6.34	4.99	4.36	3.96	4.05	3.64	3.27	2.95	3.07	2.80	3.91	4.48	9.82
AR(1)	0.82	0.85	0.87	0.88	0.88	0.90	0.91	0.92	0.99	0.95	0.86	0.72	0.41
Panel B: Correlation matrix													
CS-2	1.00												
CS-3	0.98	1.00											
CS-4	0.93	0.99	1.00										
CS-5	0.89	0.96	0.99	1.00									
FB-2	0.93	0.99	0.99	0.98	1.00								
FB-3	0.83	0.93	0.98	0.99	0.97	1.00							
FB-4	0.74	0.86	0.93	0.97	0.91	0.98	1.00						
FB-5	0.67	0.80	0.88	0.93	0.86	0.95	0.99	1.00					
PC1	0.34	0.21	0.10	0.01	0.12	-0.05	-0.17	-0.25	1.00				
PC2	0.65	0.79	0.86	0.89	0.86	0.93	0.93	0.91	-0.00	1.00			
PC3	0.33	0.28	0.21	0.14	0.29	0.13	0.01	-0.08	0.01	0.00	1.00		
CP	0.47	0.50	0.52	0.53	0.49	0.50	0.51	0.52	0.30	0.56	-0.23	1.00	
LN	0.11	0.14	0.17	0.19	0.14	0.19	0.22	0.25	-0.13	0.16	-0.14	0.20	1.00

Table IA.3: Conditioning variables

This table presents descriptive statistics for the state variables used in the empirical analysis. PMI is the Purchasing Managers' Index published by the Institute for Supply Managers and $\mathcal{U}$ is the macroeconomic uncertainty index developed in Jurado et al. (2015). The table reports mean, standard deviation, skewness, kurtosis, and first-order autocorrelation (AC(1)) of each state variable. We also report the contemporaneous correlation between the variables. The sample period is January 1962 to December 2018.

	PMI	$\mathcal{U}$
Mean	52.61	0.66
Std. dev.	6.37	0.09
Skewness	-0.61	1.63
Kurtosis	4.37	5.79
AR(1)	0.94	0.99
Correlation	-0.48	

Table IA.4: In-sample regressions

This table reports full sample least squares estimates of the slope coefficients for various linear predictive models for bond excess return. We consider five different predictors: yield spreads (Campbell and Shiller, 1991), forward spreads (Fama and Bliss, 1987), principal components of yields (Litterman and Scheinkman, 1991), the Cochrane and Piazzesi (2005) forward rate factor computed from a projection of average excess bond returns on two-, three-, four-, and five-year forward rates, and the Ludvigson and Ng (2009) macroeconomic factor computed as a projection of average excess bond returns on factors obtained from a large panel of macroeconomic variables. For each model, we report slope coefficients, Newey and West (1987) t-statistics using a bandwidth of twelve lags in parenthesis, and the adjusted R^2 in square brackets. The sample period is January 1962 to December 2018.

	2-year	3-year	4-year	5-year
	Panel A: Campbell-Shiller			
CS	2.02	2.36	2.75	3.15
	(2.67)	(2.64)	(2.85)	(3.17)
	[2.55]	[2.32]	[2.42]	[2.61]
	Panel B: Fama-Bliss			
FB	1.20	1.41	1.69	1.99
	(2.20)	(2.30)	(2.79)	(3.38)
	[1.80]	[1.68]	[1.90]	[2.14]
	Panel C: Principal components			
PC1	0.01	0.01	0.01	0.01
	(1.43)	(1.04)	(0.76)	(0.56)
PC2	0.13	0.21	0.29	0.37
	(1.72)	(2.10)	(2.46)	(2.77)
PC3	0.23	0.31	0.24	0.09
	(0.66)	(0.63)	(0.39)	(0.13)
	[1.05]	[1.09]	[1.19]	[1.30]
	Panel D: Cochrane-Piazzesi			
CP	0.65	0.88	1.11	1.36
	(4.60)	(4.30)	(4.12)	(4.08)
	[2.37]	[2.16]	[2.17]	[2.30]
	Panel E: Ludvigson-Ng			
LN	0.65	0.90	1.12	1.33
	(3.68)	(3.96)	(4.25)	(4.46)
	[6.62]	[6.47]	[6.33]	[6.15]

Table IA.5: Correlations between forecasts and macroeconomic uncertainty

This table reports correlation coefficients between out-of-sample generated forecasts from individual bond predictors (Panel A), the static (equal-weighted) forecast combination strategy (Panel B), and the dynamic forecast combination strategy (Panel C) and economic uncertainty as measured by the the macroeconomic uncertainty index ($\mathcal{U}$) from Jurado et al. (2015). We report p-values for the null of no correlation in parenthesis. The out-of-sample evaluation period runs from January 2000 to December 2018.

	2-year bond	3-year bond	4-year bond	5-year bond
	Panel A: Individual bond predictors			
CS	-0.09	-0.04	0.01	0.05
	(0.10)	(0.41)	(0.92)	(0.39)
FB	-0.04	0.07	0.12	0.15
	(0.49)	(0.21)	(0.02)	(0.00)
PC	0.03	0.04	0.05	0.06
	(0.57)	(0.5)	(0.34)	(0.23)
CP	0.12	0.11	0.10	0.10
	(0.02)	(0.04)	(0.06)	(0.07)
LN	0.44	0.46	0.47	0.48
	(0.00)	(0.00)	(0.00)	(0.00)
EH	0.43	0.38	0.34	0.32
	(0.00)	(0.00)	(0.00)	(0.00)
	Panel B: Static forecast combination			
EW	0.31	0.34	0.35	0.35
	(0.00)	(0.00)	(0.00)	(0.00)
	Panel C: Dynamic forecast combination			
PMI	0.54	0.53	0.50	0.50
	(0.00)	(0.00)	(0.00)	(0.00)
$\mathcal{U}$	0.59	0.56	0.54	0.55
	(0.00)	(0.00)	(0.00)	(0.00)
NONE	0.54	0.47	0.46	0.47
	(0.00)	(0.00)	(0.00)	(0.00)

Table IA.6: Alternative proxies for economic activity

This table reports correlation coefficients between forecasts and alternative proxies for economic activity. We use the Chicago Fed National Activity Index (Panel A), recession probabilities from Chauvet and Piger (2008) (Panel B), and log growth rates to industrial production (Panel C). We report p-values for the null of no correlation in parenthesis. The out-of-sample evaluation period runs from January 2000 to December 2018.

	2-year bond	3-year bond	4-year bond	5-year bond
Panel A: Chicago Fed National Activity Index (CFNAI)				
CS	0.10	0.04	-0.01	-0.05
	(0.07)	(0.40)	(0.89)	(0.35)
FB	0.04	-0.06	-0.13	-0.16
	(0.46)	(0.25)	(0.02)	(0.00)
PC	0.19	0.18	0.15	0.12
	(0.00)	(0.00)	(0.01)	(0.03)
CP	-0.10	-0.09	-0.08	-0.07
	(0.05)	(0.10)	(0.15)	(0.17)
LN	-0.48	-0.49	-0.50	-0.51
	(0.00)	(0.00)	(0.00)	(0.00)
EH	-0.20	-0.17	-0.16	-0.15
	(0.00)	(0.00)	(0.00)	(0.01)
EW	-0.26	-0.29	-0.30	-0.32
	(0.00)	(0.00)	(0.00)	(0.00)
PMI	-0.51	-0.55	-0.54	-0.51
	(0.00)	(0.00)	(0.00)	(0.00)
u	-0.56	-0.54	-0.54	-0.57
	(0.00)	(0.00)	(0.00)	(0.00)
NONE	-0.53	-0.49	-0.49	-0.50
	(0.00)	(0.00)	(0.00)	(0.00)
Panel B: Recession probabilities (Chauvet and Piger, 2008)				
CS	-0.01	0.02	0.05	0.08
	(0.89)	(0.72)	(0.33)	(0.13)
FB	0.03	0.09	0.14	0.16
	(0.64)	(0.09)	(0.01)	(0.00)
PC	-0.05	-0.05	-0.03	-0.01
	(0.37)	(0.35)	(0.57)	(0.86)
CP	0.10	0.08	0.06	0.05
	(0.08)	(0.16)	(0.26)	(0.32)
LN	0.56	0.57	0.58	0.59
	(0.00)	(0.00)	(0.00)	(0.00)
EH	0.18	0.13	0.11	0.09
	(0.00)	(0.01)	(0.05)	(0.09)
EW	0.37	0.38	0.38	0.38
	(0.00)	(0.00)	(0.00)	(0.00)
PMI	0.51	0.54	0.53	0.53
	(0.00)	(0.00)	(0.00)	(0.00)
u	0.55	0.56	0.54	0.55
	(0.00)	(0.00)	(0.00)	(0.00)
NONE	0.58	0.53	0.51	0.52
	(0.00)	(0.00)	(0.00)	(0.00)
Panel C: Log industrial production growth				
CS	0.07	0.06	0.03	0.01
	(0.16)	(0.28)	(0.55)	(0.86)
FB	0.07	0.01	-0.03	-0.05
	(0.16)	(0.79)	(0.61)	(0.33)
PC	0.16	0.15	0.14	0.13
	(0.00)	(0.00)	(0.01)	(0.02)
CP	-0.08	-0.07	-0.07	-0.07
	(0.16)	(0.17)	(0.18)	(0.18)
LN	-0.26	-0.27	-0.28	-0.28
	(0.00)	(0.00)	(0.00)	(0.00)
EH	-0.09	-0.10	-0.10	-0.10
	(0.11)	(0.08)	(0.07)	(0.06)
EW	-0.12	-0.14	-0.15	-0.16
	(0.03)	(0.01)	(0.01)	(0.00)
PMI	-0.23	-0.25	-0.27	-0.23
	(0.00)	(0.00)	(0.00)	(0.00)
u	-0.28	-0.25	-0.25	-0.27
	(0.00)	(0.00)	(0.00)	(0.00)
NONE	-0.25	-0.22	-0.21	-0.21
	(0.00)	(0.00)	(0.01)	(0.01)

Table IA.7: Economic Value: $\gamma = 5$

This table reports certainty equivalent return (CER) gains for various linear predictive models for bond excess return. We consider five different predictors: yield spreads (Campbell and Shiller, 1991), forward spreads (Fama and Bliss, 1987), principal components of yields (Litterman and Scheinkman, 1991), the Cochrane and Piazzesi (2005) forward rate factor, and the Ludvigson and Ng (2009) macroeconomic factor. For each model, we report the CER gains relative to the expectations hypothesis (Panels A, B, and C) and a static (equal-weighted) forecast combination strategy (Panel D). PMI denotes the Purchasing Managers Index published by the Institute for Supply Management and $\mathcal{U}$ is the macroeconomic uncertainty index from Jurado et al. (2015). CER gains are based on an investor with mean-variance preferences and a relative risk aversion of $\gamma = 5$. The out-of-sample evaluation period runs from January 2000 to December 2018.

	2-year	3-year	4-year	5-year
	Panel A: Individual bond predictors against EH			
CS	-0.91	-0.88	-0.52	-0.25
	(0.94)	(0.87)	(0.74)	(0.62)
FB	-0.62	-0.68	-0.55	-0.34
	(0.88)	(0.86)	(0.80)	(0.67)
PC	-2.06	-2.46	-2.41	-2.36
	(0.99)	(0.96)	(0.93)	(0.9)
CP	-0.80	-1.20	-1.31	-1.36
	(0.96)	(0.94)	(0.91)	(0.87)
LN	0.61	1.39	2.41	3.24
	(0.01)	(0.01)	(0.00)	(0.00)
	Panel B: Static forecast combination against EH			
EW	0.03	0.25	0.70	1.08
	(0.46)	(0.32)	(0.13)	(0.07)
	Panel C: Dynamic forecast combination against EH			
PMI	0.28	0.59	1.07	1.47
	(0.19)	(0.14)	(0.05)	(0.02)
$\mathcal{U}$	0.19	0.53	1.22	1.60
	(0.27)	(0.14)	(0.02)	(0.01)
NONE	0.12	0.30	0.76	1.07
	(0.34)	(0.28)	(0.10)	(0.07)
	Panel D: Dynamic forecast combination against EW			
PMI	0.25	0.34	0.37	0.39
	(0.02)	(0.04)	(0.05)	(0.03)
$\mathcal{U}$	0.16	0.28	0.52	0.52
	(0.08)	(0.02)	(0.00)	(0.02)
NONE	0.09	0.04	0.06	-0.01
	(0.15)	(0.35)	(0.28)	(0.52)

Table IA.8: Sharpe ratio improvements

This table reports certainty Sharpe ratio improvements for various linear predictive models for bond excess return. We consider five different predictors: yield spreads (Campbell and Shiller, 1991), forward spreads (Fama and Bliss, 1987), principal components of yields (Litterman and Scheinkman, 1991), the Cochrane and Piazzesi (2005) forward rate factor, and the Ludvigson and Ng (2009) macroeconomic factor. For each model, we report the Sharpe ratio improvement relative to that earned by a buy-and-hold investor. PMI denotes the Purchasing Managers Index published by the Institute for Supply Management and $\mathcal{U}$ is the macroeconomic uncertainty index from Jurado et al. (2015). If $R^2_{OS} < 0$ we do not report a value. The out-of-sample evaluation period runs from January 2000 to December 2018.

	2-year	3-year	4-year	5-year
SR	0.90	0.83	0.78	0.74
		Panel A: Individual bond predictors		
CS	-	-	0.05	0.11
FB	-	0.10	0.13	0.14
PC	-	-	-	-
CP	-	-	-	-
LN	-	-	-	-
		Panel B: Static forecast combination		
EW	0.38	0.35	0.34	0.33
		Panel C: Dynamic forecast combination		
PMI	0.48	0.37	0.35	0.43
$\mathcal{U}$	0.59	0.44	0.41	0.36
NONE	0.41	0.35	0.36	0.35

Figure IA.1: Empirical power curves

This figure shows the rejection frequency (empirical power) of the multivariate test for equal predictive ability with a nominal size of 5% and data-generating process given by (IA.B.20) with the first element in $\boldsymbol{\mu}$ deviating and the remaining elements are set to zero. The first element of $\boldsymbol{\mu}$ is set to $c\hat{\eta}$ where $\hat{\eta}$ is the average absolute loss differentials across all models within the low and high economic activity states defined in the empirical section and $c \in [0, 2.5]$. We use 10,000 Monte Carlo replications. We implement a conditional test that sets $\boldsymbol{h}_t = (1, \tilde{h}_t)'$, and use three samples sizes referred to as short (120 observations), medium (348 observations) and long (1,000 observations). The left (right) panel depicts results where $\boldsymbol{\varepsilon}_{t+1}$ in (IA.B.20) is sampled from the empirical loss differentials when forecasting the 2-year (5-year) bond. The value of p indicates the dimension of the test arising from the number of comparing models less one.
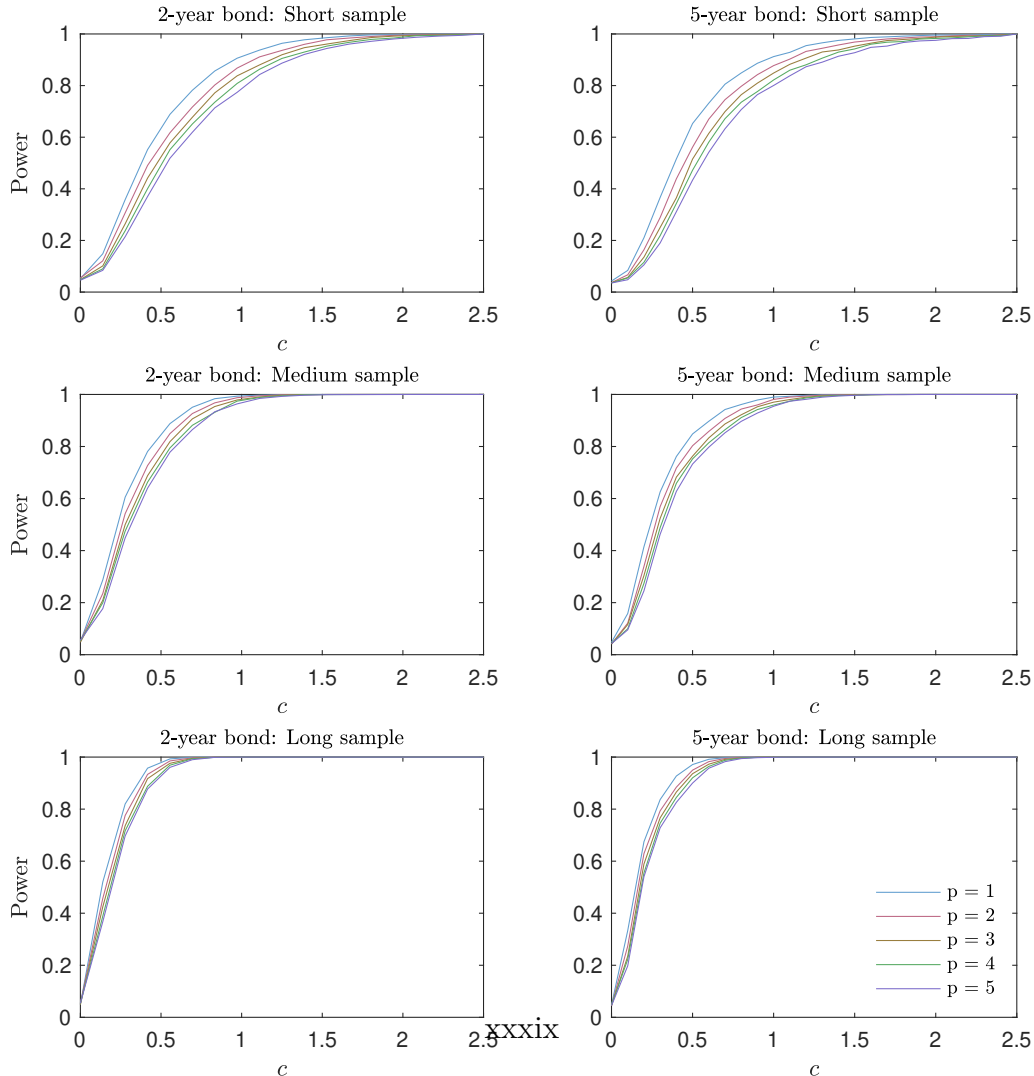
Figure IA.2: Inclusion plots across states

This figure displays the inclusion of each predictive model into the best set of models. Green (yellow) shaded ares represent periods of high (low) states of the Purchasing Managers' Index (PMI) (left) and the Jurado et al. (2015) macroeconomic uncertainty index ($\mathcal{U}$) (right) identified using the 20% and 80% quantiles of the series. White areas are normal times. We consider five different predictors: yield spreads (Campbell and Shiller, 1991), forward spreads (Fama and Bliss, 1987), principal components of yields (Litterman and Scheinkman, 1991), the Cochrane and Piazzesi (2005) forward rate factor, and the Ludvigson and Ng (2009) macroeconomic factor. EH denotes the benchmark expectations hypothesis model. Inclusion of a predictive model is marked with +. The out-of-sample evaluation periods runs from January 2000 to December 2018.
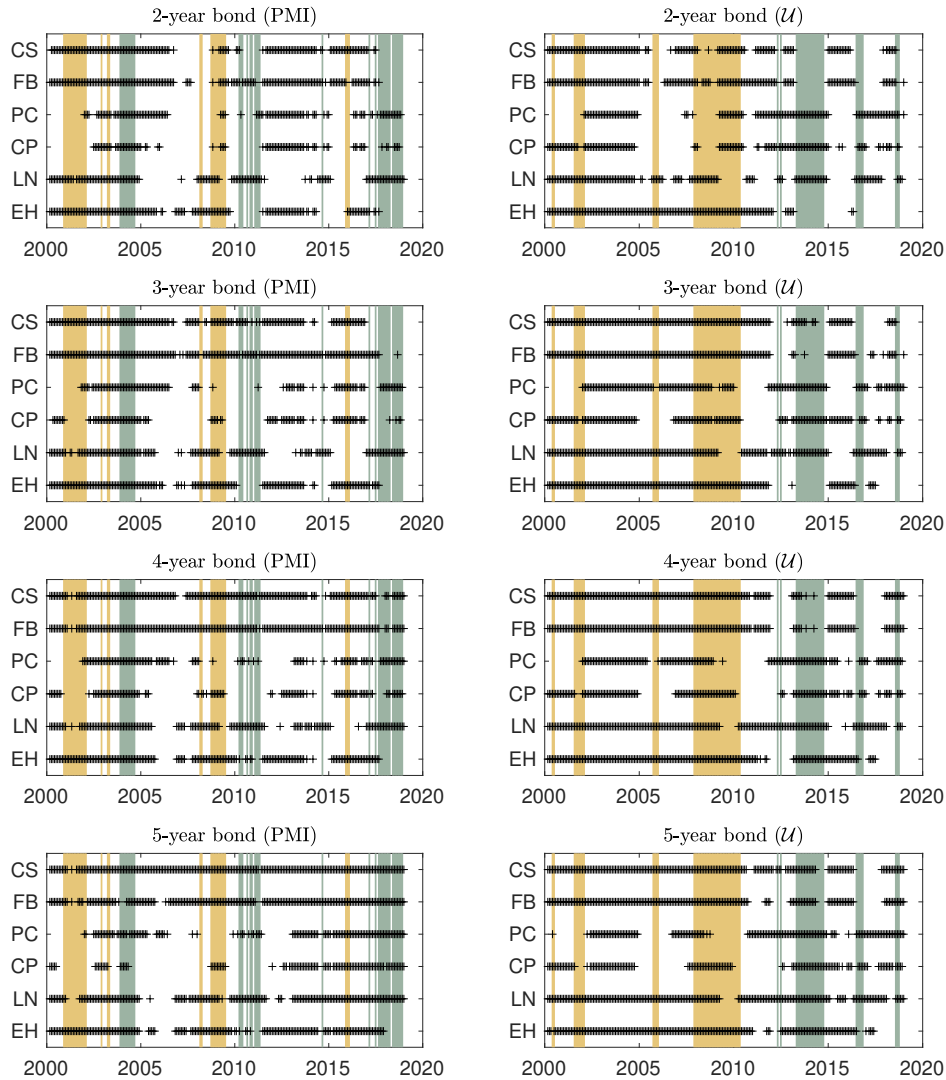

xl

Figure IA.3: Size of the set of best models

This figure illustrates the size of the set of best predictive models for each of the four bond maturities and conditioning variables. Green (yellow) shaded ares represent periods of high (low) activity and uncertainty, respectively, where activity is measured using the Purchasing Manager's Index (PMI) published by the Institute for Supply Management and uncertainty is the macroeconomic uncertainty index ($\mathcal{U}$) proposed in Jurado et al. (2015). High (low) episodes are identified using the 80% (20%) quantiles of their time series. White areas are normal times. The out-of-sample evaluation periods runs from January 2000 to December 2018.

Purchasing Managers Index (PMI)

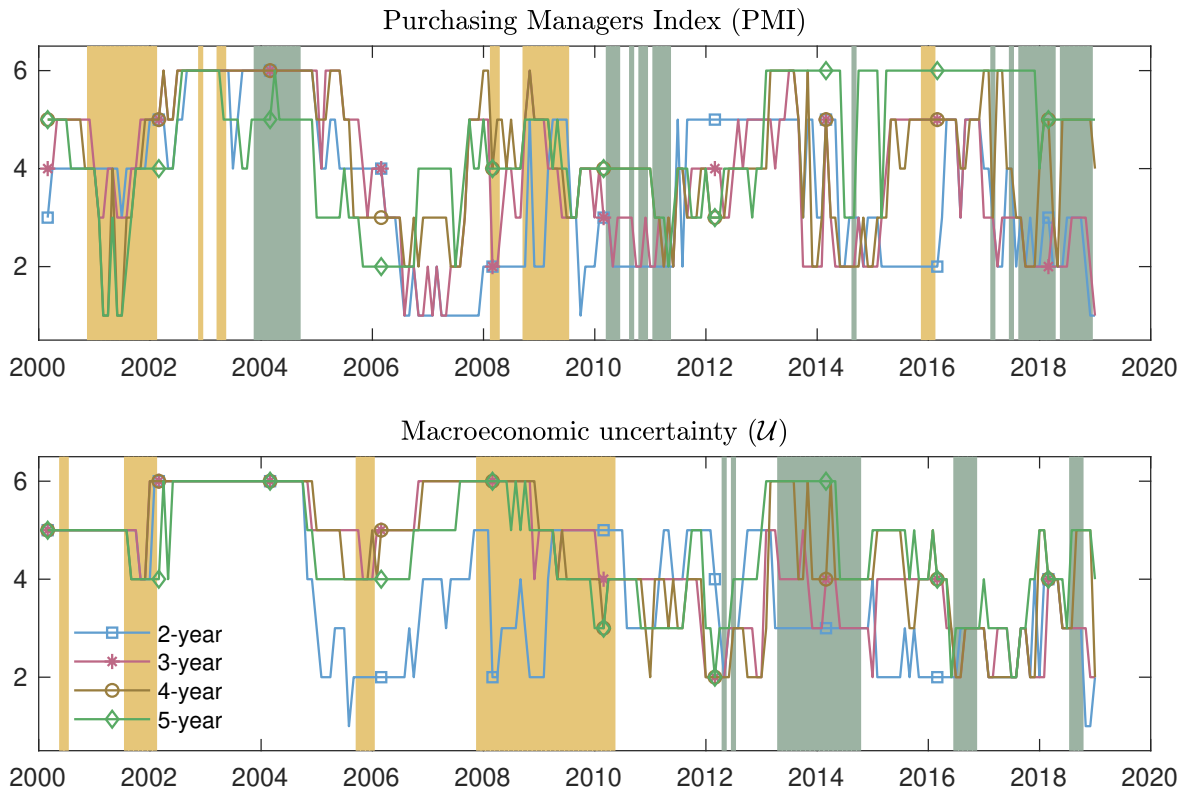Macroeconomic uncertainty ($\mathcal{U}$)

Figure IA.4: Full sample elimination order

This figure displays the full sample elimination order of predictive model in high, normal, and low states separately for the Purchasing Managers' Index (PMI) (left graphs) and the macroeconomic uncertainty index ($\mathcal{U}$) of Jurado et al. (2015) (right graphs) using the 20% and 80% quantiles of their time series. White squares denote models included in the best set of models and numbered tiles denotes eliminated models and their elimination order. We consider five different predictors: yield spreads (Campbell and Shiller, 1991), forward spreads (Fama and Bliss, 1987), principal components of yields (Litterman and Scheinkman, 1991), the Cochrane and Piazzesi (2005) forward rate factor, and the Ludvigson and Ng (2009) macroeconomic factor. The out-of-sample evaluation periods runs from January 2000 to December 2018.

High activity (Models × Maturity)

Models	2	3	4	5
EH	1	1	2	2
LN				
CP	2	2	1	1
PC				
FB	3			
CS	4			

Low uncertainty (Models × Maturity)

Models	2	3	4	5
EH	3	2	3	4
LN				
CP	1	1	1	1
PC				
FB	4		2	2
CS	2		4	3

Normal activity (Models × Maturity)

Models	2	3	4	5
EH	5	3		
LN	3			
CP	2	2		
PC	1	1	1	1
FB				
CS	4			

Normal uncertainty (Models × Maturity)

Models	2	3	4	5
EH	4			
LN				
CP	1	1	1	1
PC	2	2	2	2
FB				
CS	3			

Low activity (Models × Maturity)

Models	2	3	4	5
EH				
LN	1	1	1	1
CP	4	3		
PC	2	2	2	2
FB	5			
CS	3			

High uncertainty (Models × Maturity)

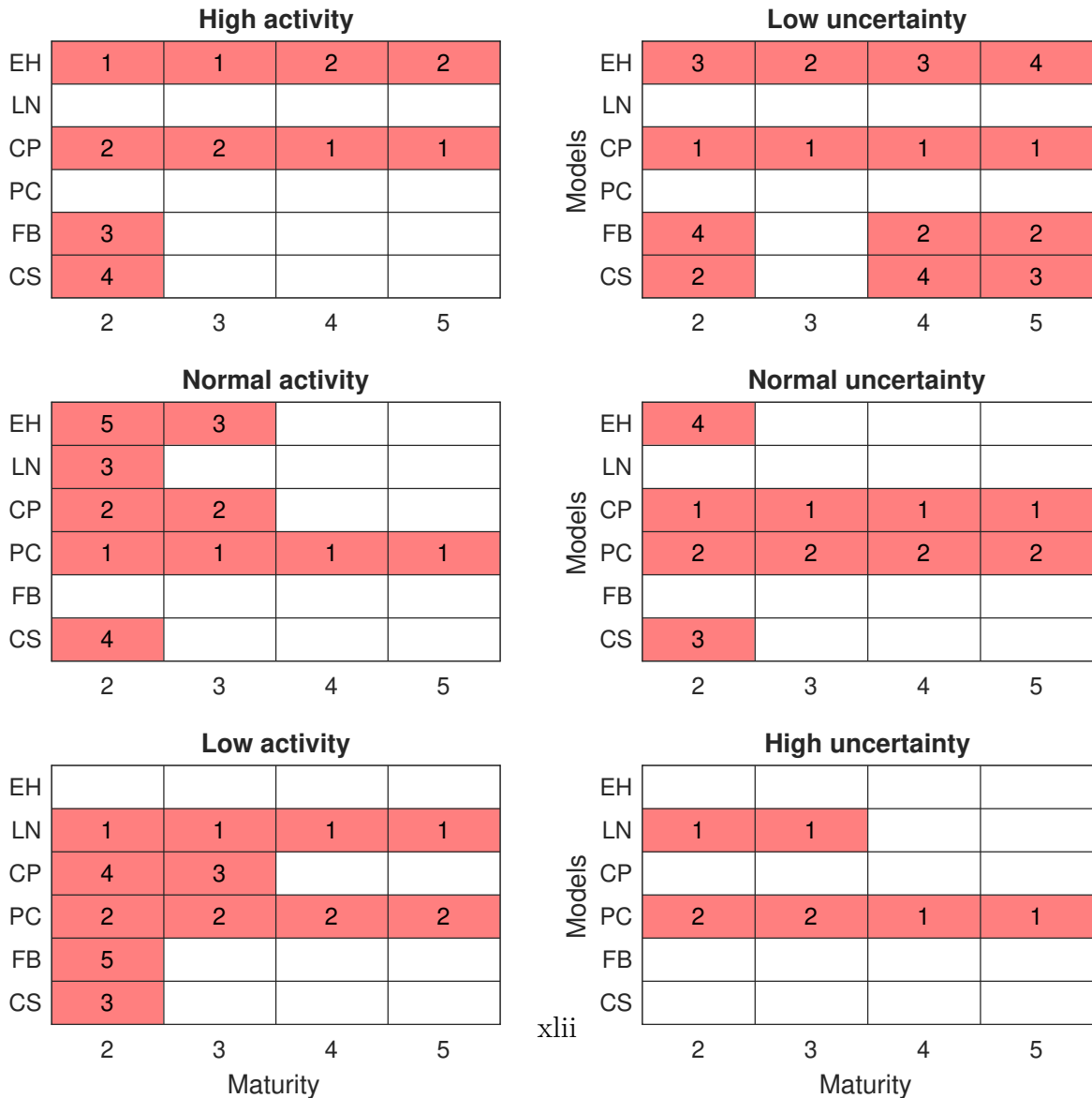| Models	2	3	4	5
EH				
LN	1	1		
CP				
PC	2	2	1	1
FB				
CS				

xlii

Figure IA.5: Relative certainty equivalent returns

This figure plots the recursively updated cumulative difference in realized utility from the EH benchmark model and the ith predictor model over the out-of-sample evaluation period. We consider five different predictors: yield spreads (Campbell and Shiller, 1991), forward spreads (Fama and Bliss, 1987), principal components of yields (Litterman and Scheinkman, 1991), the Cochrane and Piazzesi (2005) forward rate factor, and the Ludvigson and Ng (2009) macroeconomic factor. We also consider a simple equal-weighted combination of the individual forecasts. A positive (negative) slope indicates that the predictive model delivers more (less) utility than the EH benchmark. Green (yellow) shaded ares represent periods of high (low) activity and uncertainty, respectively, where activity is measured using the Purchasing Managers' Index (PMI) (left column) and uncertainty (right column) is the index developed by Jurado et al. (2015). High (low) episodes are identified using the 80% (20%) quantiles of their time series. White areas are normal times. The out-of-sample evaluation period runs from January 2000 to December 2018.