Asymmetric Violations of the Spanning Hypothesis*

Gustavo Freire†[1] and **Raul Riva**‡[2]

[1]Econometric Institute, Erasmus School of Economics
[2]Finance Department, Northwestern University

Click here for latest version

Abstract

We document that the Spanning Hypothesis, which most macro-finance term structure models imply, is violated asymmetrically along the U.S. nominal yield curve. Using an interpretable reduced-form representation of yields and different Machine Learning techniques, we find that macroeconomic variables enhance the predictability of yields only for the shorter end of the yield curve, with no evidence of improvements for the longer end. Such asymmetry leads to higher predictability of bond returns for shorter maturities and economic gains to a mean-variance investor, adding nuance to the debate about yield curve predictability. We provide evidence that this extra predictability comes from more accurate predictions of the path of short rates and not from the term premia. We further show that moments in which the Federal Reserve deviates from the Taylor Rule are associated with sharper increases in the predictability of the short end when conditioning on macro data.

Keywords: Bond risk premia, yield curve, Spanning Hypothesis, Nelson-Siegel factors, machine learning

JEL Classification: G11, G12, G17

*We are especially grateful to Caio Almeida, Torben Andersen, Robert Korajczyk, and Viktor Todorov for their guidance and insights. Joshua Mollner provided important feedback on several sections. We also thank Cassiano Alves, Philip Barret, Michael Bauer, Anthony Diercks, Diogo Duarte, Marcelo Fernandes, Andrei Gonçalves, Felipe Iachan, Leonardo Iania, João Victor Issler, Sebastian Jaimungal, Onno Kleen, Robin Lumsdaine, Marcelo Medeiros, Geert Mesters, Eduardo Mendes, Humberto Moreira, Ioana Neamtu, Ioannis Papantonis, Filip Obradovic̀, Matthew Pritsker, Miguel Santana, André Santos, Rosnel Sessinou, Costis Skiadas, Rodrigo Targino, Dennis Umlamdt, Amílcar Velez, Michel van der Wel, and Cynthia Wu for important comments and suggestions. We are indebted to participants from the 2023 SoFiE Summer School in Brussels, the XXIII Meeting of the Brazilian Finance Society, the 2023 Brazilian School of Time Series Econometrics and Statistics, the 2023 Trends in Macroeconometrics Conference at UIUC, the 17th Brunel Conference on Macro and Financial Econometrics, COPPEAD-UFRJ Finance Seminar, the Northwestern Kellogg Brown Bag Series, the FinEML Seminar Series, the 2024 Machine Learning in Finance Bootcamp at the University of Toronto, the 2024 Midwest Macro Conference, the 2024 SoFiE Annual Meeting, the 2024 QFFE Conference in Marseille, the 2024 IAAE Annual Conference in Thessaloniki, the 2024 Meeting of the Bachelier Finance Society, and the 2024 European Economic Association meeting for helpful suggestions.

†freire@ese.eur.nl.

‡raul.riva@kellogg.northwestern.edu.

1 Introduction

The nominal U.S. yield curve, typically regarded as the risk-free rate for different maturities across international markets, is of central importance for traders, long-term capital managers, households, and policymakers. The workhorse-type models researchers deploy to characterize the joint dynamics of yields are the so-called arbitrage-free Dynamic Term Structure Models (DTSMs). Such models usually have two defining features: underlying risk factors evolving as Markovian processes and a flexible parametrization of a stochastic discount factor that, through a no-arbitrage condition, delivers bond prices and yields. Hence, these models are essentially characterized by the mapping from risk factors to yields they generate.[1]

One pervasive feature of most DTSMs is the *Spanning Hypothesis*. This property entails that the mapping from risk factors to yields can be inverted, i.e., the dynamics of risk factors can be uncovered from the dynamics of yields. Hence, an observer whose information set is comprised only of the yield curve itself would have the exact same information as another observer who knows only the current realization of risk factors. Despite its name, the Spanning Hypothesis is not a hypothesis in the context of these models – it is an implication. In other words, it arises naturally from their theoretical structure, and it is part of the predictions these models have for how interest rates should behave, both in the cross-section and over time. In this paper, we concentrate on the implications of the Spanning Hypothesis for the conditional first moment of yields.

The ability to invert the factors-to-yields mapping has strong empirical implications. For instance, if the Spanning Hypothesis is valid, any agent forecasting interest rates, which is a critical step for portfolio allocation and risk management, should lean towards models with few variables since the American yield curve is well-described by a low-rank factor structure (Litterman and Scheinkman, 1991). Similarly, market makers should be able to hedge interest-rate derivatives, one of the most liquid types of contracts, using few assets (Collin-Dufresne and Goldstein, 2002). Moreover, there is empirical evidence that the types of risk that indeed command a premium in models with and without this invertibility property are very different (Adrian et al., 2013). Even if such property is violated by the data, an important question is how far we are from this benchmark.

The focus of this paper is testing an implication of the Spanning Hypothesis for the conditional expectation of interest rates and bond returns: given today's yield curve, no other state variable should provide additional forecasting power when one targets either future yields or bond returns. This is equivalent to saying that the current yield curve is all one needs to predict the future yield curve at any point in time. The state variables we entertain in this paper consist of a large panel of macroeconomic signals about the U.S. economy. We show that the Spanning Hypothesis is violated asymmetrically across the maturity spectrum: given information already contained in the yield curve itself, macroeconomic signals provide larger increases in predictability for the shorter end of the yield curve. In fact, we are not able to reject the Spanning Hypothesis for longer maturities.

We bring nuance to the current debate about the Spanning Hypothesis and its implication for yield

[1] See Duffee (2013a,b) for an extensive review.

curve predictability. Previous papers such as Ludvigson and Ng (2009), Cooper and Priestley (2009), Greenwood and Vayanos (2014), Joslin et al. (2014), and Cieslak and Povala (2015) claim that different variables are able to predict bond returns after already controlling for the information in the yield curve. The typical approach in this literature is running predictive linear regressions of bond returns on the current yields plus additional regressors and evaluating the statistical significance of these additional associated coefficients. Bauer and Hamilton (2018) show that this evidence is fragile, highlighting that small samples of persistent regressors are challenging terrain for standard inference in linear regressions.[2] Bauer and Rudebusch (2017) provide simulation results indicating that such regression evidence is not necessarily inconsistent with the Spanning Hypothesis.

We depart from the usual methodology in three important ways. First, we do not focus on in-sample results; instead, we concentrate on (pseudo-) out-of-sample forecasts. This is closer to the real task a professional analyst would have at hand over time and partially guards us against the benefit of hindsight the econometrician always has. Second, we do not focus on a specific variable to enlarge the information set generated by the current yield curve. We tie our hands and use a large monthly panel of 126 macroeconomic variables for the American economy that is commonly used in empirical macroeconomic research - the FRED-MD dataset from the St. Louis Fed, described in McCracken and Ng (2016).[3]

The choice of working with a data-rich environment leads to a methodological problem due to the high dimensionality of the data, as we cannot simply run linear regressions. Our third point of departure from much of the previous literature is deploying different techniques from the Machine Learning literature such as Lasso, Ridge, Elastic Net, Random Forest, and regression on principal components to document that violations of the Spanning Hypothesis by variables from the FRED-MD universe are asymmetric across maturities, in the sense that they are stronger at shorter maturities. These techniques were developed precisely to handle forecasting problems when there are many possible signals to extract information from. We rely on these techniques to solve the type of problem they were created for, taking the prediction task as a way of testing a salient implication from a large class of DTSMs.

Our first exercise sets up a recursive forecasting problem of 1-year bond holding returns. At each point in time t, we predict the returns for holding bonds of different maturities that will be realized 12 months in the future, both with and without the aid of principal components extracted from the FRED-MD variables *up to* time t. In the next month, after acquiring one more data point, we extract the principal components once again and repeat the exercise. After computing the ratio between the mean squared error (MSE) in prediction with and without the macroeconomic signals, we show that it increases with the maturity of the initial bond, indicating that the macroeconomic signals are more helpful in modeling the dynamics of the short end of the curve than the long end.

[2]Out of the mentioned papers, only the evidence from Cieslak and Povala (2015) survives the more stringent inference procedure proposed by Bauer and Hamilton (2018). However, that paper does not speak directly to the asymmetry we focus on in our work.

[3]We believe this dataset is, on the one hand, comprehensive enough to let the Spanning Hypothesis be tested in a data-rich environment but, on the other hand, also a standard choice in the macroeconomic and financial forecasting literature. Naturally, we are silent about *other* variables that might or might not generate violations of the Spanning Hypothesis but are not included in our chosen dataset. Our methodology can, nevertheless, be applied to any other balanced dataset.

In order to assess the economic significance of this finding, we study the investment problem of a mean-variance investor who chooses between trading risky bonds and investing in the risk-free rate. This investor is able to improve the Sharpe ratio of her optimal strategy by up to 0.2 annualized points by taking advantage of violations of the Spanning Hypothesis. These improvements are, however, concentrated on shorter maturities - which is consistent with the idea that violations of the Spanning Hypothesis happen asymmetrically across maturities.

The only challenge when forecasting returns is forecasting future zero-coupon yields, by definition. Aiming at studying the behavior of yields along the maturity dimension, we adopt the Nelson-Siegel (Nelson and Siegel, 1987) representation of the yield curve. It is a reduced-form model that fits the American yield curve remarkably well and decomposes yields of different maturities into combinations of a long-run factor, a short-run factor, and a medium-run factor. We prove that if yields follow the Nelson-Siegel representation, then bond returns can be written essentially as a linear combination of innovations of these factors.

We use different Machine Learning techniques to show that the macroeconomic variables provide valuable information for, and only for, the short-run Nelson-Siegel factor. These methods, when exploiting information contained both in the yield curve and in the FRED-MD dataset, can generate forecasts that are more precise than a simple random walk and a baseline model that extracts information only from the yield curve itself. This short-run factor is empirically close to the slope of the yield curve. Hence, our results can be interpreted as macroeconomic data being important to predict the slope of the yield curve and not other factors. Our decomposition of bond returns explicitly shows how the predictability of this short-run factor gets distributed along different maturities.

Taking a step to dissect where predictability comes from, we use the standard accounting identity for zero-coupon yields, which implies that, at any point in time, the yields of different maturities can be decomposed into two components: one component related to the expected path of short rates in the future and a term premium component. The exact separation between the two of them requires a model to discipline expectations.[4] We adopt the term structure model from Adrian et al. (2013), which is the default one used by the New York Fed to generate their real-time term premium measurements.[5] We feed the time series of expected short rates and term premia generated by their model into our methodology and find that macroeconomic variables from the FRED-MD dataset help predicting the path of short rates. Additionally, we find that the same asymmetry we document when predicting bond returns and the Nelson-Siegel factors manifests itself when predicting the path of short rates. In contrast, we find no evidence that the macroeconomic variables we use improve the forecast of the term premium after considering the information already in the yield curve.

The future path of short rates is heavily influenced by monetary policy both because the Federal Reserve controls the Fed Funds rate and because of its usual communication about the *path* of the policy

[4]For example, in a world where all agents are risk-neutral, current yields should reflect only the path of expected short rates. The term premium should be zero in this situation. This special case is also the baseline for the so-called *Expectations Hypothesis*, which implies that current yields purely reflect the expected path of short rates. See Gürkaynak and Wright (2012) for a review.

[5]See https://www.newyorkfed.org/research/data_indicators/term-premia-tabs#/interactive.

rate.[6] Aside from its mandate of targeting low inflation and maximum employment, there is ample evidence that monetary policy responds to more variables than those two quantities (Bernanke and Boivin, 2003; Moench, 2008). Additionally, not even professional forecasters fully understand how policy will be conducted given the current business cycle.[7] In most DTSMs, however, the short rate usually is a *known* function of the underlying risk factors. Such a function typically has no time-varying components and is fully understood by the representative agent whose stochastic discount factor prices bonds. This is equivalent to saying that monetary policy follows a simple, static, fully-understood rule. Therefore, there is no scope, within these models, for any uncertainty regarding monetary policy or changes in the interpretation of the current business cycle by the monetary authority.

Recent evidence by Schmeling et al. (2022) shows that returns from opening positions on Fed Funds futures and interest rate swaps tend to be concentrated in moments in which the Federal Reserve deviates from a dynamically updated traditional Taylor Rule. Professional forecasters also get less precise as the deviations become larger. These deviations can be broadly understood as indication that there is some new element, or some new assessment of the business cycle, guiding monetary policy at that moment in time. Based on this evidence and our previous results, we conjecture that the macroeconomic signals we entertain help predict the shorter end of the yield curve because they might be informative about monetary policy. Under this conjecture, part of the asymmetry we document would stem from the asymmetric force the Federal Reserve exerts over the yield curve.

To test this conjecture, we fit a traditional Taylor Rule comprised of inflation and unemployment to the Fed Funds data and track deviations from it. We show that forecasts for the short-run Nelson-Siegel factor computed with information both from the yield curve and from the macroeconomic data get *more precise* as deviations from the Taylor Rule increase. Such correlation does not exist when we study the precision of forecasts done *only* with information from the yield curve itself. For the long-run factor, we find no correlation between precision and deviations from the Taylor Rule, no matter the information set used. This result lends credibility to our initial conjecture that business-cycle conditions may be particularly useful for anticipating monetary policy, which is translated to higher statistical precision when forecasting shorter rates with an enlarged information set.

The remainder of the paper is organized as follows. After briefly reviewing the related literature, Section 2 describes the data we use. Section 3 studies the forecast of bond risk premia over different maturities and analyzes its economic significance. Section 4 provides a novel decomposition of bond excess returns in terms of innovations of Nelson-Siegel factors. Section 5 contains the main empirical results for the prediction of the short-, medium- and long-run factors using yield curve data and other macroeconomic state variables. Section 6 shows that the help provided by macroeconomic data comes from the expected path of short rate and not from term premia. Section 7 tests our conjecture about deviations from the Taylor Rule. Finally, Sections 8 and 9 further discuss our results and conclude the paper, respectively.

[6] See, for example, Gürkaynak et al. (2005), Kuttner (2018), Swanson (2023), and Swanson and Jayawickrema (2024) for the empirical effects of central bank communication on yields.

[7] See Piazzesi et al. (2015) for model-based evidence of how forecasters consistently overestimate the persistence of the slope of the yield curve and Cieslak (2018) for survey-based evidence on how professional forecasters have been consistently surprised over time by the Federal Reserve.

Related Literature

Our paper relates to the vast literature studying the term structure of interest rates. First, we relate to papers forecasting interest rates and bond risk premia. Fama and Bliss (1987), Campbell and Shiller (1991), and Cochrane and Piazzesi (2005) study bond return predictability using information only from the yield curve using in-sample regressions. Providing evidence against the Spanning Hypothesis instead, several papers document the predictive power of macroeconomic variables for bond returns (Cooper and Priestley, 2009; Ludvigson and Ng, 2009; Joslin et al., 2014; Greenwood and Vayanos, 2014; Cieslak and Povala, 2015), but also focus on in-sample exercises. Duffee (2011) and Bauer and Rudebusch (2017) argue that measurement error can lead to violations of the Spanning Hypothesis, while Bauer and Hamilton (2018) show that small-sample distortions weaken this earlier evidence from in-sample predictive regressions.

More recently, Bauer and Rudebusch (2020) and Favero et al. (2023) show that deviations of yields from time-varying long-run trends contain predictive power for future bond returns. Bianchi et al. (2021) demonstrate that non-linearities captured by Machine Learning methods provide stronger evidence in favor of out-of-sample bond return predictability, but their work is silent about the type of asymmetry we document. Borup et al. (2023) document that bond return predictability is state-dependent and related to the current level of economic activity, but they investigate fewer maturities than us and focus primarily on forecasts using only information already in the yield curve. Huang and Shi (2023) use regularization techniques to build a business-cycle factor from the FRED-MD dataset and show that it provides extra predictive power when predicting excess bond returns, but they largely focus on in-sample exercises and do not speak to the asymmetry we concentrate on. Bauer and Chernov (2024) show that option-implied conditional yield skewness, which is by definition a forward-looking measure, has predictive power for bond returns and relate that finding to biased beliefs from investors. They are also silent about the asymmetry we document because they focus on predicting an average measure of bond returns across maturities.

Our first main contribution is to provide new out-of-sample evidence that violations of the Spanning Hypothesis are asymmetric in the dimension of yield maturity. We stress that our goal is not to draw a comprehensive comparison across different forecasting methodologies like Gu et al. (2020) (equity return prediction), Bianchi et al. (2021) (bond return prediction), or Medeiros et al. (2021) (inflation prediction). Instead, we show that our main empirical finding holds regardless of the forecasting method we use. Given a specific forecasting method, we are interested in assessing whether enlarging the econometrician's information set with macroeconomic variables will contribute to better predictability.

Second, we are related to papers using dynamic versions of the Nelson-Siegel model (Nelson and Siegel, 1987) as in Diebold and Rudebusch (2013). Diebold and Li (2006) and van Dijk et al. (2013) specify autoregressive models for Nelson-Siegel factors to forecast the yield curve. Moench (2008) performs a similar exercise but augments these systems with principal components from the FRED-MD dataset. Similarly, Altavilla et al. (2014), Altavilla et al. (2017) and Fernandes and Vieira (2019) augment the Nelson-Siegel model with forward-looking variables, while Coroneo et al. (2016), Guidolin and Pedio (2019) and Caldeira et al. (2023) introduce unspanned macroeconomic factors, regime-switching

and stochastic volatility in this framework, respectively. Hännikäinen (2017) examines the predictive power of Nelson-Siegel factors for future industrial production. In contrast to these papers, we investigate whether macroeconomic variables help predict future realizations of the Nelson-Siegel factors after controlling for the current yield curve in an environment where these factors are estimated only with information already in the yield curve. Their time series evolution is not constrained a priori, and they are estimated only with information from a single date at each point in time. We contribute to this literature by showing that all the predictability business-cycle-type macroeconomic data can provide is concentrated on the short-run Nelson-Siegel factor.

Third, we make contact with the literature that investigates the economic value of bond return predictability. Thornton and Valente (2012) provide evidence that the economic significance of violations of the Expectations Hypothesis, which implies that bond risk premia is constant, is weak. Sarno et al. (2016) show this depends on macroeconomic uncertainty. Gargano et al. (2019) reconcile this evidence with the statistical rejection of the Expectations Hypothesis, which is common in the literature, with a model incorporating stochastic volatility and unspanned macroeconomic factors. Tackling a different problem, we show that violations of the Spanning Hypothesis for short-maturity bonds translate into economic gains for an investor with mean-variance preferences, while the same is not necessarily true for long-maturity bonds. None of the previous papers document the asymmetry in the Sharpe ratio improvement we find.

Finally, we are also related to more recent literature, which recognizes that the central bank's reaction function is unknown and might change over time. Cieslak (2018) shows that the Fed has consistently surprised professional forecasters. The same type of phenomenon is documented by Schmeling et al. (2022), who further show that both forecast errors from professional forecasters and returns of Fed Funds futures positions are correlated with deviations from a Taylor Rule. Bauer and Swanson (2023) show that monetary policy shocks, typically identified using high-frequency data, can have different interpretations when one accounts for an unknown central bank reaction function. Bauer et al. (2024) use survey data to document how different perceptions about this reaction function evolve over time. We contribute to this literature by showing that macroeconomic signals about the business cycle help the econometrician better predict short rates when the Federal Reserve deviates from a Taylor Rule similar to the one in Schmeling et al. (2022).

2 Data

We rely on three data sets for our empirical exercises. First, our yield curve data comes from Liu and Wu (2021). Second, our data on macroeconomic variables is taken from FRED-MD, a monthly data set maintained by the St. Louis Federal Reserve Bank and described in McCracken and Ng (2016). Third, from the New York Fed website, we acquire time series for the expected path of short rates and term premia, as estimated in Adrian et al. (2013).

The yield curve data set we use represents an improvement over other commonly used sources of yield curve data for the U.S., namely data sets constructed under either the methodology from Fama and Bliss (1987) or Gurkaynak et al. (2007). Concerning the former, Liu and Wu (2021) provide information

about longer maturities since the Fama and Bliss (1987) data set currently stored on CRSP files covers yields only up to five years. This is crucial for us since we will contrast the behavior of the short end of the yield curve *vis-à-vis* the long end.

On the other hand, the yield curve proposed by Gurkaynak et al. (2007) is known to generate relatively high pricing errors for bonds of short maturity while completely ignoring the very short end of the yield curve. Liu and Wu (2021) show that their methodology, which uses a kernel-based smoothing technique, reduces the pricing errors across different maturities in comparison to Gurkaynak et al. (2007), who work with a reduced-form factor model for the interpolation of yields.

Although the yield curve from Liu and Wu (2021) is available at daily frequency, we use end-of-month data since that is the highest frequency we can work with if we want to match it with macroeconomic data. For most of our analysis, we pick all maturities from one to ten years covering the period 1973-2021, which implies we work with a balanced panel of zero-coupon yields.[8] We start the sample in 1973 since the 10-year bond started being traded in late 1972.

Regarding macroeconomic indicators, we rely on the FRED-MD data set as described in McCracken and Ng (2016). This database is maintained by specialists at the St. Louis Fed that take care of data anomalies that might occur when bundling information from different sources and is freely available online. Our version of this data set consists of 126 monthly macroeconomic series classified into eight distinct groups by the specialists: prices, labor market, housing, interest and exchange rates, monetary aggregates and credit measures, output measures, orders and inventories, and stock-market-related measures.

These variables are typically in levels and might not be stationary as reported. We apply simple transformations to make the data stationary, following the recommendations from McCracken and Ng (2016).[9] Table D.1 in Appendix D reports the complete list of variables used and the transformations applied to them, together with a short description of what they are and their respective FRED code. Figure A.1, in Appendix A, reports the spectral decomposition of the sub-sample from the FRED-MD data set we use. The first principal component explains roughly a quarter of the total variation of the data set, while the first three principal components command around 40% of the total variation.

This data set has, in different forms, been used in forecasting exercises whenever researchers need a standardized and freely available "data-rich" environment. An earlier version has been used in seminal work from Stock and Watson (2002a) and Stock and Watson (2002b), for example. More recently, Ludvigson and Ng (2009) and Bianchi et al. (2021) use it to forecast excess bond returns, while Medeiros et al. (2021) use it in an inflation forecasting exercise. We use a balanced panel of macroeconomic indicators.[10]

[8] The data set provided by Liu and Wu (2021) includes maturities of up to 30 years (360 months) after the introduction of the 30-year securities during the 1980s. However, the liquidity of these longer-term bonds over time has been disputed. Hence, we adopt a similar strategy as Bianchi et al. (2021) in their investigation and focus on maturities up to 120 months.

[9] Further details are available at https://research.stlouisfed.org/econ/mccracken/fred-databases/.

[10] To guard against the fact that macroeconomic data might be released with some delay, we lag all macroeconomic series by an extra month in our empirical implementations so we alleviate look-ahead biases. Our (unreported) robustness checks show that this has no material impact in empirical results.

3 Bond Risk Premia

3.1 Forecasting

We start by investigating the predictability of holding bond returns in excess of the risk-free rate. Throughout the paper, we concentrate on a holding period of one year, although our data is at the monthly frequency. Hence, we work with 12-steps-ahead forecasts, which is a standard framework in this literature (Cochrane and Piazzesi, 2005; Ludvigson and Ng, 2009; Joslin et al., 2014; Bianchi et al., 2021).[11]

More specifically, we let $y_t^{(n)}$ be the n-year zero-coupon yield at month t.[12] Then, $y_t^{(1)}$ represents the 1-year risk-free rate at time t. We denote by $xr_{t+12}(n)$ the excess return over the 1-year risk-free rate obtained from the purchase of an n-year bond at time t and its subsequent sale at time $t+12$:

$$xr_{t+12}(n) = n \cdot y_t^{(n)} - (n-1) \cdot y_{t+12}^{(n-1)} - y_t^{(1)}. \tag{1}$$

As it is obvious from our notation, this return is only known at time $t+12$. The only random term, conditional on the information up to time t, is the second one. From that point of view, variables that help forecast $xr_{t+12}(n)$ should also help forecast $y_{t+12}^{(n-1)}$.

Under the Spanning Hypothesis, conditional on the information summarized by the yield curve at time t, no state variable should be able to enhance the forecast of $xr_{t+12}(n)$ and, equivalently, $y_{t+12}^{(n-1)}$. Importantly, this should hold regardless of the maturity n. Macroeconomic indicators are natural state variables to test the spanning hypothesis as many models used by macroeconomists and financial economists tie together aggregate variables and the dynamics of interest rates (see, e.g., the discussion in Bauer and Rudebusch, 2017).

Since the number of macroeconomic variables in our data set is large, it is not feasible to add all of them to a linear model for $xr_{t+12}(n)$ and estimate it by ordinary least squares (OLS), for example. Using only a few variables would also force us to pick from a large menu. However, under the Spanning Hypothesis, they should all be irrelevant for forecasting the excess bond returns. In the same spirit as Ludvigson and Ng (2009), we use principal components of the FRED-MD data set to summarize macroeconomic information.[13]

Let PC_t denote a $K \times 1$ vector of principal components extracted from the FRED-MD data set while C_t is a $d \times 1$ vector summarizing information from the yield curve. We study the following predictive regression:

$$xr_{t+12}(n) = \alpha_n + \theta_n' C_t + \gamma_n' PC_t + \epsilon_{t+12,n}. \tag{2}$$

There are different reasonable choices for C_t. One can adopt a strategy similar to Cochrane and Piazzesi (2005) and let $C_t = (y_t^{(1)}, \mathbf{f}_t')'$, where $\mathbf{f}_t$ stacks a sequence of forward rates implied by the yield curve at time t.[14] Another strategy, building on Litterman and Scheinkman (1991), would be considering the first

[11] See recent discussions about this environment of overlapping returns in Bauer and Hamilton (2018) and Feng et al. (2022).

[12] That implies that the log-price of an n-year bond that has a \$1 face-value at time t is $p_t = -n \cdot y_t^{(n)}$.

[13] McCracken and Ng (2016) find that the entire data set is well described by six to eight principal components.

[14] The forward rate for maturity n at time t is defined as $f_t^{(n)} = n \cdot y_t^{(n)} - (n-1) \cdot y_t^{(n-1)}$.

three principal components of the yield curve itself since much of the total variation can be explained by this low-rank factor structure.

In this framework, the usual approach in the literature to test the spanning hypothesis is to run regression (2) over the whole sample and test whether $\gamma_n = 0$. However, inference based on these in-sample predictive regressions is challenging due to severe small-sample distortions (Bauer and Hamilton, 2018). In contrast, we focus directly on out-of-sample risk premia forecasts, denoted by $\widehat{xr}_{t+12}(n)$, and assess whether allowing for $\gamma_n \neq 0$ results in better predictive ability.

We estimate a linear model as in (2) with an expanding window, keeping track of the one-year-ahead forecasts. We start our out-of-sample period in January 1990 following Bianchi et al. (2021). For concreteness, the forecast for January 1990 is made with all data available up to January 1989. We take principal components of the macroeconomic variables available up to January 1989 and use them as regressors for the forecasting exercise. After fitting the model with the available data, we use the estimated parameters to forecast the returns that will be realized in January 1990. As we move ahead in time, the amount of data used both in extracting principal components of the FRED-MD data set and estimating equation (2) increases. We repeat this exercise under the validity of the spanning hypothesis ($\gamma_n = 0$) and under alternative specifications when we vary the number of included principal components. In total, we have 384 out-of-sample forecasts. While we focus on out-of-sample exercises, we also report in-sample results in Appendix B, which are in line with the out-of-sample evidence presented below.

Each set of predictions generates a time series of squared prediction errors. Under the Spanning Hypothesis, a model with macroeconomic data should display no better performance than a model that imposes $\gamma_n = 0$. To assess the enhancement provided by the addition of macroeconomic variables to our forecasting scheme, we compute the ratio of the mean squared errors:

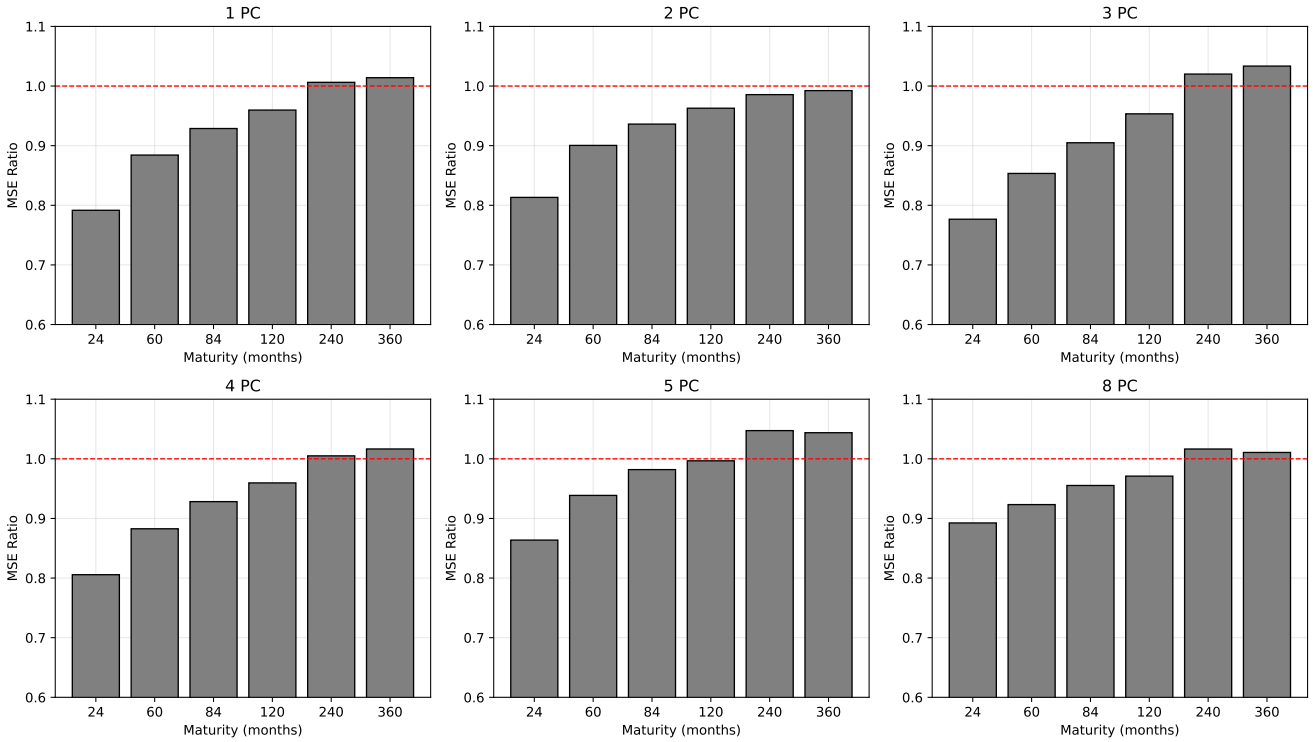
$$\text{MSE Ratio} = \frac{\sum_{t=t_0}^{T} (xr_t(n) - \widehat{xr}_t(n))^2}{\sum_{t=t_0}^{T} (xr_t(n) - \widehat{xr}_{t|\gamma_n=0}(n))^2}, \tag{3}$$

where $\widehat{xr}_t(n)$ is a forecast made at $t-12$ from a model that allows $\gamma_n \neq 0$ while $\widehat{xr}_{t|\gamma_n=0}(n)$ is the analogous forecast under the Spanning Hypothesis.

Figure 1 reports the statistic defined in (3) for different maturities and different numbers of principal components from our macroeconomic data. Under the Spanning Hypothesis, the bars from Figure 1 should oscillate around unity. However, the shorter maturities display MSE ratios generally much lower than one, which implies that conditioning on macroeconomic variables is helpful in forecasting excess bond returns. For instance, for the 24-month maturity, we document a decrease of about 20% of the mean squared error. In contrast, as the maturity increases, the ratio approaches unity and sometimes goes slightly above. This suggests that violations of the spanning hypothesis are stronger at the shorter end of the yield curve. For instance, the reduction in out-of-sample MSE is never greater than 10% for the 10-year maturity.

For this exercise, our preferred method for spanning the yield curve is using the forward rates. The reason is that forward rates are directly available from the zero-coupon yields and require no estimation.

Figure 1: Relative MSE predicting returns using forward rates as control. For each maturity, we show the ratio between the MSE attained with different numbers of principal components from the macroeconomic data and the baseline model that uses information only from the yield curve itself. The sample for maturity of less than 120 months ranges from 1973 to 2021, while it starts in 1985 for the other maturities. For any of the maturities, the out-of-sample period starts in January 1990. We use the linear model in (2) to make the forecasts.



Nonetheless, as a robustness check, we repeat the out-of-sample forecasting exercise, letting C_t stack the first three principal components of the yield curve, which is extracted sequentially as done with the macroeconomic data. Results are reported in Figure A.2 in Appendix A. The same asymmetry arises, where the absolute decrease in the MSE is even larger for shorter maturities.[15]

Figure A.3 in Appendix A reports results for the same exercise depicted in Figure 1, but now using real-time data. The same general pattern arises. In order to deal with revisions and missing data when taking principal components, we adopt the EM algorithm as in Stock and Watson (2002b). We conclude that the overall evidence from Figure 1 cannot be explained by data revisions.

We also test whether the reductions in the MSEs presented in Figure 1 are statistically significant. Under the Spanning Hypothesis, differences in predictive ability should be attributed to sampling noise. For each of the maturities and specifications considered, we use the methodology from Diebold and Mariano (1995) to assess the significance of our results.[16] We let C_t stack the forward rates. The p-values

[15]This is likely due to imprecise estimation of the principal component of yields. If this extraction is noisy, C_t will not provide enough information about the yield curve at time t. If information spanned by the true yield curve and not captured by a noisy C_t is present in the macroeconomic principal components, relaxing the restriction of $\gamma_n = 0$ will have a disproportionately powerful effect in reducing the MSE. This would make one more likely to reject the Spanning Hypothesis across maturities just because the econometrician does not have all the information available in yields in the first place. We prefer to err on the side of caution and use the forward rates stacked in C_t since they require no estimation.

[16]We allow for autocorrelation in the forecasting errors and deploy the HAC estimator from Newey and West (1987) to

for the test are reported in Table 1. The null hypothesis is of equal predictive ability, which would be implied by the Spanning Hypothesis.

Table 1: p-values (Diebold and Mariano, 1995) for testing whether macro data enhances forecasting using forward rates as controls. See the discussion in the caption of Figure 1. Variances for the Diebold and Mariano (1995) test are computed using the HAC estimator of Newey and West (1987).

	Maturity in months					
	24	60	84	120	240	360
1 PC	0.00	0.01	0.02	0.05	0.74	0.92
2 PC	0.00	0.01	0.01	0.04	0.16	0.32
3 PC	0.02	0.01	0.04	0.13	0.81	0.96
4 PC	0.04	0.06	0.13	0.24	0.55	0.65
5 PC	0.18	0.28	0.42	0.48	0.80	0.84
6 PC	0.21	0.25	0.35	0.38	0.69	0.66
7 PC	0.16	0.09	0.13	0.16	0.34	0.28
8 PC	0.24	0.23	0.32	0.37	0.59	0.57
9 PC	0.12	0.11	0.19	0.33	0.75	0.80
10 PC	0.15	0.12	0.19	0.28	0.79	0.51

A few patterns appear from these p-values. First, at the 5% level, we reject the spanning hypothesis for maturities of up to 7 years using one, two or three principal components from the macroeconomic data. Second, given any number of principal components from the FRED-MD data set, the p-values generally increase with maturity. This is consistent with the idea that violations of the spanning hypothesis are stronger at the shorter end of the curve. In fact, at the 5% level, we can only reject the null for the 10-year maturity once, and we can never reject the null for the 20-year and the 30-year maturities at usual levels. Third, the p-values typically increase when a larger number of principal components is considered. This is intuitive: adding principal components implies estimating more coefficients stacked in γ_n, which increases estimation uncertainty. Since we are using a quadratic loss function to evaluate our forecasts, there is a bias-variance trade-off. Larger models, understood as models that consider a larger number of principal components, might overfit in-sample and generate poor out-of-sample forecasts, making it hard to reject the spanning hypothesis in that case.

3.2 Economic Significance

Now, we assess whether these violations are economically meaningful. We present a setup in which a consumer who builds portfolios of bonds based on our methodology is able to achieve a higher Sharpe ratio using macroeconomic data. This improvement, however, is stronger when trading shorter maturity bonds than when trading bonds with longer maturities, exactly as one would expect in face of our previous statistical results. Our environment is similar to the one in Thornton and Valente (2012).

We study the problem of a consumer who takes monthly trading decisions and holds for a full year whatever bond portfolio she has assembled at time t. The consumer has mean-variance utility over gross returns of her portfolio $R_{p,t}(\boldsymbol{w}_{t-12}) = 1 + y^{(1)}_{t-12} + \boldsymbol{w}'_{t-12}\boldsymbol{x}\boldsymbol{r}_t$, where $\boldsymbol{w}_{t-12}$ is a vector of portfolio

compute the variance required by Diebold and Mariano (1995).

weights chosen at $t-12$ and $\boldsymbol{xr}_t$ is a vector of risk premia known at time t, which collects the excess returns for bonds with different maturities. At each point in time, the consumer solves the following optimization problem:

$$\max_{\boldsymbol{w}_t} \left\{ \mathbb{E}_t \left[R_{p,t+12}(\boldsymbol{w}_t) \right] - \frac{A}{2} \cdot \text{Var}_t \left[R_{p,t+12}(\boldsymbol{w}_t) \right] \right\},$$

for some aversion coefficient $A > 0$.

We further define $\boldsymbol{\mu}_{t+12|t} \equiv \mathbb{E}_t \left[\boldsymbol{xr}_{t+12} \right]$ and $\Sigma_{t+12|t} \equiv \mathbb{E}_t \left[\left(\boldsymbol{xr}_{t+12} - \boldsymbol{\mu}_{t+12|t} \right) \left(\boldsymbol{xr}_{t+12} - \boldsymbol{\mu}_{t+12|t} \right)' \right]$ as the conditional risk premia and the conditional covariance matrix of these excess returns, respectively. Given these two objects, we know the solution to the problem above in closed form:

$$\boldsymbol{w}_t^* = \frac{1}{A} \cdot \Sigma_{t+12|t}^{-1} \boldsymbol{\mu}_{t+12|t}.$$

Our methodology readily delivers estimates of $\boldsymbol{\mu}_{t+12|t}$ with and without macroeconomic data. We do not have, nonetheless, a full model for the covariance matrix. The Spanning Hypothesis also restricts the time-series evolution of conditional second moments with important implications for hedging securities, but this is out of the scope of this paper.[17] Instead, we follow Thornton and Valente (2012) and Bianchi et al. (2021) and use a non-parametric estimator:

$$\widehat{\Sigma}_{t+12|t} \equiv \sum_{i=0}^{\infty} \widehat{\boldsymbol{\epsilon}}_{t-i} \widehat{\boldsymbol{\epsilon}}_{t-i}' \odot \Omega_{t-i}, \quad \Omega_{t-i} = \alpha \cdot e^{-\alpha \cdot i} \mathbf{1}\mathbf{1}',$$

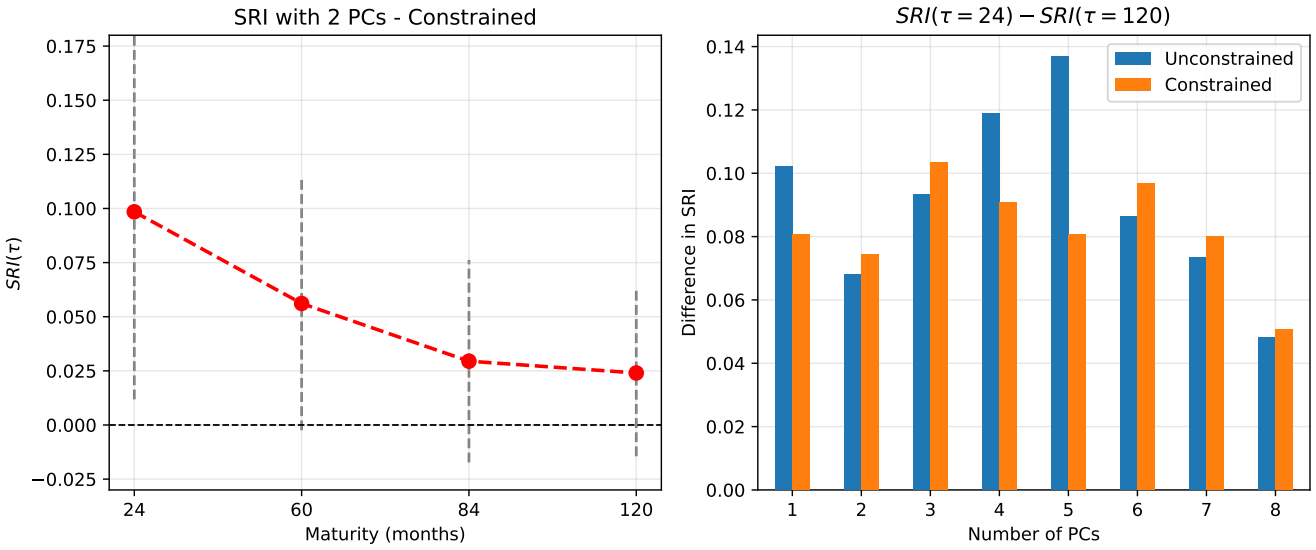
where $\widehat{\boldsymbol{\epsilon}}_{t-i}$ is an observed forecast error realized at $t-i$, $\alpha > 0$ is a scalar, $\mathbf{1}$ is a vector of ones, and $\odot$ is the element-by-element multiplication operator. This estimator exponentially downweights the past and gives more importance to recent forecast errors. The decay speed is controlled by α. We choose $\alpha = 0.05$ as in Bianchi et al. (2021). We can compute this estimator using forecast errors obtained with and without macroeconomic data. Under the Spanning Hypothesis, these forecast errors should be the same in population, and the mean-variance investor should get no benefit by conditioning her decisions on state variables that are not the yield curve itself.

Since we are interested in the economic significance of violations of the Spanning Hypothesis for each maturity, we focus on a univariate allocation exercise where the mean-variance investor can trade only the risk-free rate and one risky bond at a time. More specifically, for each maturity n, we have that the optimal allocation on the risky bond is $w_t^{(n)} = \mu_{t+12|t}^{(n)} / A \cdot \sigma_{t+12|t}^{(n)}$, where $\mu_{t+12|t}^{(n)}$ is the bond risk premia forecast for a given model and $\sigma_{t+12|t}^{(n)}$ is the diagonal element of $\widehat{\Sigma}_{t+12|t}$ corresponding to the bond of maturity n. We can perform this exercise using different numbers of principal components, effectively creating the trading counterpart of the exercise showcased by Figure 1.

We keep track of the investor's decisions and keep track of the realizations of $R_{p,t}$. The metric we use to assess the performance of this trading strategy is the Sharpe ratio, defined as the ratio between the average realized risk premium of the bond portfolio and its standard deviation. We keep the out-of-sample period the same (1990-2021) and use $A = 3$.

[17] See, for example, Collin-Dufresne and Goldstein (2002); Li and Zhao (2006, 2009); Andersen and Benzoni (2010); Backwell (2021), and Riva (2024) for a discussion of the connection between the Spanning Hypothesis and the volatility of yields.

Figure 2: Left: improvement in Sharpe Ratio when trading each maturity with vs without macroeconomic signals. Dotted lines are 95% confidence intervals following Ledoit and Wolf (2008). The out-of-sample period is 1990-2021. We impose short-sale constraints and allow for no leverage. Right: Difference in the Sharpe ratio improvement between the 2-year and the 10-year maturity. Blue bars refer to the unconstrained case (short sales and leverage are allowed), while the orange ones refer to the constrained one.



Our setup abstracts away from trading costs related to leverage and short positions in some of these bonds. Aiming at bringing the setup closer to what a practitioner could do, we study the problem under two different realistic scenarios. The first one is a strict scenario where we further constrain $0 \leq \boldsymbol{w}_t \leq 1$, i.e., when the consumer cannot short any bond and cannot take on any leverage. Taking long positions with no borrowing in these assets should be as close to costless as possible since it is a large, liquid market.

The second scenario is less restrictive and allows the consumer both to short bonds and to take on leverage. However, we still want to rule out extreme positions. We follow Welch and Goyal (2007) and Ferreira and Santa-Clara (2011) and impose $-1 \leq \boldsymbol{w}_t \leq 2$. We compute the Sharpe ratio of this trading strategy, with and without macro data, and focus on the Sharpe ratio improvement across different maturities τ, denoted by $SRI(\tau)$:

$$SRI(\tau) \equiv SR^{(\text{Macro})}(\tau) - SR^{(\text{SH})}(\tau).$$

The baseline Sharpe ratio achieved by the consumer is between 0.2 and 0.3 using only information from the yield curve, no matter how we control for it. The left panel from Figure 2 plots $SRI(\tau)$ for four different maturities under our more restrictive scenario. In that case, we rely on information from the FRED-MD variables using two principal components. We also plot 95% confidence intervals based on Ledoit and Wolf (2008).

For the shortest maturity (2 years), we have a statistically significant improvement in Sharpe ratio of about 0.1 annualized points. As we increase the maturity, these improvements decrease, which follows directly from the asymmetry documented by Figure 1. For the 7- and 10-year maturities, these improvements are not statistically significant anymore. This qualitative behavior remains true as we

change the number of principal components from the macroeconomic signals we use. In Appendix A, Figure A.4 shows the decreasing profile for $SRI(\tau)$ for different numbers of principal components, keeping the no-short sale constraint fixed. Figure A.5 repeats the exercise allowing for short sales and leverage when trading bonds and shows the same qualitative behavior.[18]

The panel on the right shows the difference between $SRI(\tau = 24)$ and $SRI(\tau = 120)$ across different numbers of principal components and across both of our scenarios. For both groups of bars and across the number of principal components, we find that this difference is positive. That implies a higher improvement when trading with macroeconomic signals at the shorter end of the yield curve in contrast to the smaller improvements for the longer end. Intuitively, more precise predictions of bond returns lead to improvements in Sharpe ratios, however these improvements are also asymmetric along the maturity dimension.

4 Decomposing Bond Risk Premia

We now develop a methodology that will enable us to investigate, interpret, and quantify this finding more thoroughly. The first step in our approach is modeling the entire yield curve with a parsimonious reduced-form model that fits the U.S. yield curve well. We adopt a model in the spirit of Nelson and Siegel (1987), Diebold and Li (2006) and Diebold et al. (2006). For a certain maturity τ measured in months, we assume that the zero-coupon yield at time t follows:

$$y_t^{(\tau)} = \beta_{1,t} + \beta_{2,t}\left(\frac{1-e^{-\lambda_t\tau}}{\lambda_t\tau}\right) + \beta_{3,t}\left(\frac{1-e^{-\lambda_t\tau}}{\lambda_t\tau} - e^{-\lambda_t\tau}\right), \tag{4}$$

where $(\beta_{1,t}, \beta_{2,t}, \beta_{3,t}, \lambda_t) \in \mathbb{R}^4$ are random variables. This model is widely used by central banks and practitioners due to its simplicity and flexibility.[19] At any point in time, yields are a linear combination of three factors. The weights carried by each of these factors, however, depend on the specific maturity τ. The positive scalar λ_t is called the decay parameter since it affects how fast the loadings change across maturities.

These factors have been interpreted as the level, the slope, and the curvature of the yield curve, respectively. In fact, Diebold and Li (2006) and more recently Hännikäinen (2017) show that they have very high correlations with empirical counterparts of the actual level, slope, and curvature of the yield curve. Our preferred interpretation is closely linked to but slightly different from this traditional interpretation.

We interpret β_1 as a long-run factor since $\lim_{\tau\to\infty} y_t^{(\tau)} = \beta_{1,t}$. The interpretation that it represents the level of the yield curve builds on the idea that changes in β_1 move all yields together by the same amount. On the other hand, we interpret β_2 as a short-run factor. The loading on β_2, for a fixed positive value of λ_t, starts at 1 and monotonically converges towards zero as τ increases. Hence, changes in β_2 will affect the shorter end of the yield curve disproportionately more than the longer end, all else equal. The parameter λ_t controls how fast this decay happens. The interpretation of β_2 as the slope of the yield

[18] The confidence intervals get wider as we increase the number of principal components, but this is expected as we are estimating more coefficients from the same amount of data.

[19] See the discussion in Almeida and Vicente (2008) for example. We follow the parametrization of Diebold and Li (2006).

curve stems from the fact that:

$$\lim_{\tau\to\infty} y_t^{(\tau)} - \lim_{\tau\to 0} y_t^{(\tau)} = -\beta_{2,t}. \tag{5}$$

Finally, we take β_3 as a medium-run factor. Its loading starts at zero and also converges towards zero as τ increases, but it attains an interior maximum. Therefore, it will affect neither the very short end of the yield curve nor the very long end, concentrating its effect on intermediate maturities. The precise location of this maximum is also affected by λ_t. The fact that the loading on β_3 has a hump-shaped format motivates calling it the curvature factor.

Aside from its flexibility in fitting the yield curve, this reduced-form model offers a convenient way of isolating the short, medium and long ends of the yield curve, which is crucial for our methodology. It also offers a number of other advantages when compared to other methods:

1. We have precise interpretations of the factors themselves by construction.
2. Principal component analysis, when used as a way to identify factors in an approximate factor structure setting, suffers from an identification problem. Factors and loadings, in that case, are identified only up to a rotation. In principle, there is no *a priori* best possible rotation. The Nelson and Siegel (1987) approach solves this identification problem by assuming a parametric form of loadings.
3. The implied price at time t of a bond that pays one dollar τ months ahead is given by $P_t(\tau) = e^{-\tau y_t^{(\tau)}}$. This is also called the discount curve when seen as a function of τ. The Nelson and Siegel (1987) method ensures that the discount curve starts at one and converges to zero, as it is implied by virtually all economic models. This does not need to be the case if we use, for example, splines-based methodologies.

Perhaps more importantly, the parametric form of the loadings in equation (4) allows us to develop a decomposition of the excess bond returns. From now on, we assume a constant $\lambda_t = \lambda > 0$ since that will be part of our estimation strategy, which we discuss later. Our decomposition is given in the proposition below.

Proposition 1. *Suppose the yield curve follows* (4) *and assume that the decay parameter is a positive constant* $\lambda_t = \lambda > 0$. *Define* $\theta \equiv 12\lambda$. *Then, the one-year excess bond return for a maturity of n years is given by:*

$$\begin{aligned} xr_{t+12}(n) = (n-1)&\left[\beta_{1,t} - \beta_{1,t+12}\right] \\ &+ \left(\frac{1-e^{-\theta(n-1)}}{\theta}\right)\left[e^{-\theta}\beta_{2,t} - \beta_{2,t+12}\right] \\ &+ \left(\frac{1-e^{-\theta(n-1)}}{\theta} - ne^{-\theta(n-1)} + 1\right)\left[e^{-\theta}\beta_{3,t} - \beta_{3,t+12}\right] + \left(1-e^{-\theta(n-1)}\right)\beta_{3,t+12}. \end{aligned} \tag{6}$$

Proof. See Appendix E. □

This proposition shows that, for any maturity, the excess bond returns can be written as combinations of the innovations on the factors. The terms in the parentheses, for a given $\lambda > 0$, are not random and depend only on the maturity n. For long maturities, the term preceding innovations in the long-run factor β_1 is dominant since it increases linearly with maturity. Conversely, the term preceding innovations in the short-term factor β_2 is bounded above by $1/\theta$ and becomes relatively less important as the maturity increases. A similar phenomenon happens with the loading multiplying the innovations in the medium-run factor since it is bounded above by $1 + 1/\theta$. Finally, the very last term displays the future level of the medium-run factor multiplied by a nonrandom loading which is close to zero for shorter maturities and bounded above by 1 as the maturity increases.

The decomposition from Proposition 1 indicates that the predictability of the excess bond returns must be tied to the predictability of the factors on the right-hand side. Since the contribution of each component of the decomposition above depends on the maturity, being able to perfectly predict, for example, the long-run factor should impact the predictability of excess bond returns at longer maturities but should not be as relevant for the shorter ones. By the same token, improved predictability of the short-run factor should be translated to improved predictability of excess bond returns of shorter maturities without too much impact for the longer maturities. This fact is useful for us since it suggests a natural way to test the asymmetry in the violations of the Spanning Hypothesis across different regions of the yield curve.

4.1 Estimation

We now turn to the estimation of the model in (4). Different estimation procedures have been used in the literature. We adopt the OLS approach from Diebold and Li (2006) due to its numerical stability and simplicity. This method has also been more recently advocated by van Dijk et al. (2013), Diebold and Rudebusch (2013) and Hännikäinen (2017).

Given any constant value $\lambda > 0$ for the decay parameter, an OLS regression of the cross-section of zero-coupon yields on the loadings is able to identify the factors. One cross-sectional regression is required for each time t. This effectively implies that we impose no restriction on the dynamics of the factors between date t and any other date t' since separate linear regressions are estimated for different dates. This provides great flexibility to fit the yield curve month by month. It is also computationally simple and stable since the estimators for the factors are known in closed form. We analyze the period 1973-2021 and use all yields available from 1 to 120 months in the data set provided by Liu and Wu (2021).[20]

[20] We have conducted (unreported) robustness checks regarding using even longer yields up to 360 months whenever available and also using only a few fixed maturities as in Diebold and Li (2006). We found that the time series for the factors were essentially indistinguishable from the ones based on our approach. Given λ, there are only three parameters to estimate, and the behavior of these factors is diverse enough that a few yields in the cross-section are enough to identify them. Extra results are available upon request.

Formally, our estimator for the factors at time t is given by:

$$\begin{bmatrix}\beta_{1,t}\\ \beta_{2,t}\\ \beta_{3,t}\end{bmatrix} = \left(M'M\right)^{-1}M'Y_t, \qquad M = \begin{bmatrix}1 & \left(\frac{1-e^{-\lambda\tau_1}}{\lambda\tau_1}\right) & \left(\frac{1-e^{-\lambda\tau_1}}{\lambda\tau_1} - e^{-\lambda\tau_1}\right)\\ \vdots & \vdots & \vdots\\ 1 & \left(\frac{1-e^{-\lambda\tau_N}}{\lambda\tau_N}\right) & \left(\frac{1-e^{-\lambda\tau_N}}{\lambda\tau_N} - e^{-\lambda\tau_N}\right)\end{bmatrix}, \quad Y_t = \begin{bmatrix}y_t^{(\tau_1)}\\ \vdots\\ y_t^{(\tau_N)}\end{bmatrix}, \tag{7}$$

where $N = 120$ is the cross-sectional size. We pick $\lambda = 0.0609$ as in Diebold and Li (2006) as the decay parameter. This value implies that the maximum effect of the medium-run factor is attained at the 30-month horizon. Additionally, it facilitates comparisons with other studies that followed the same methodology.

Figure 3: Estimated factors using the OLS approach of Diebold and Li (2006), with $\lambda = 0.0609$. For each date, factors are estimated by running a linear regression of observed yields on the loadings. We use all yields from 1 to 120 months that are available from the data set provided by Liu and Wu (2021). The sample ranges from January 1973 to December 2021.

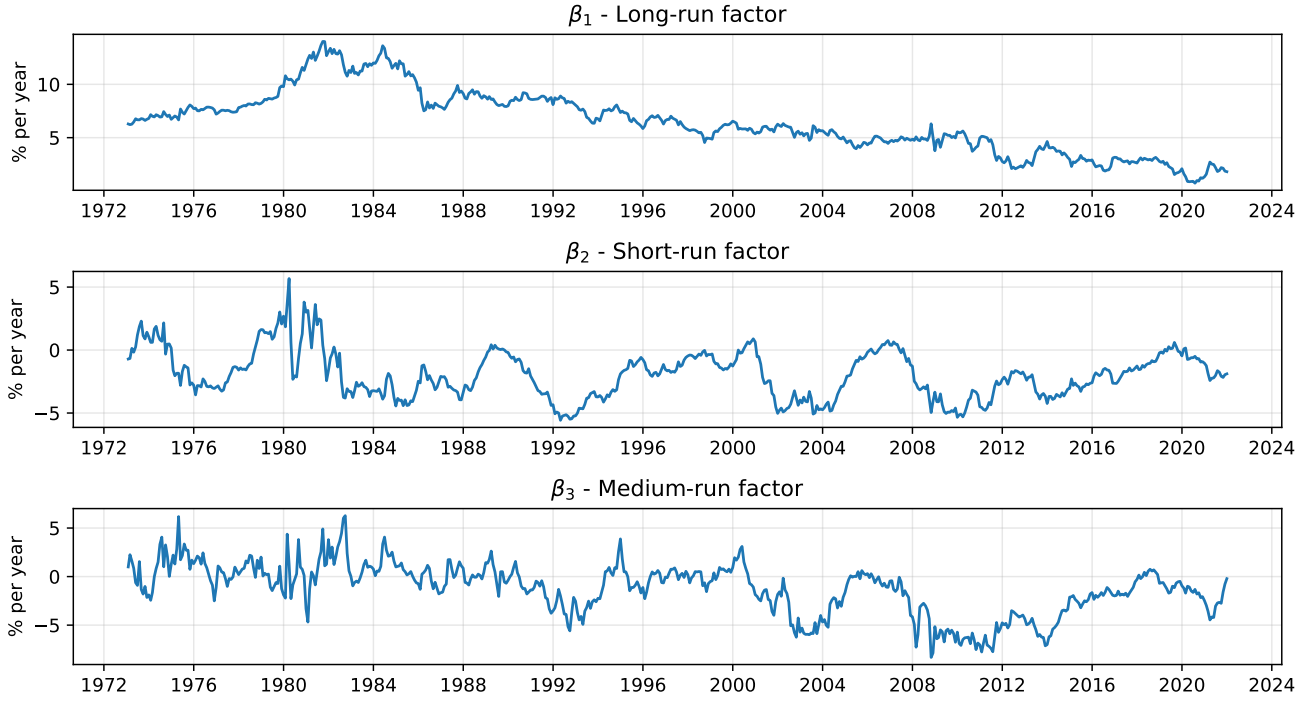


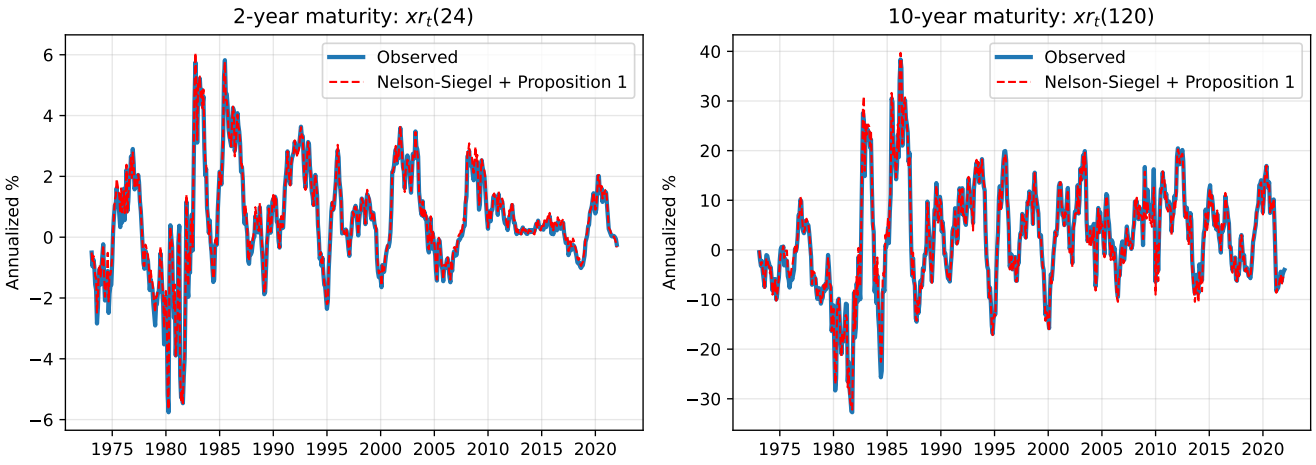
Figure 3 displays the estimated time series for each factor. It is immediate to see that they are persistent time series, which is not surprising as the cross-section of zero-coupon yields is persistent as well. The scale of the long-run factor is also slightly greater than the realizations of the other two. The long-run factor is always positive, while the other two oscillate around zero. Table 2 shows summary statistics for these factors, and the p-values of an augmented Dickey and Fuller (1979) test in two versions: conditioning on only a constant and conditioning also on a linear time-trend. At the 5% level, we reject the hypothesis of an unit root both for β_2 and β_3, for both versions of the test. For the long-run factor, however, we only reject the null at the 10% level when we also control for a linear time trend. In general, the long-run factor is much more persistent than the other ones.

We have highlighted the advantages of the Nelson-Siegel approach, but the question of whether the model is indeed able to fit the yield curve data is purely empirical. Importantly, given Proposition

Table 2: Summary statistics for the estimated Nelson-Siegel factors. For each date, factors are estimated by running a linear regression of observed yields on the loadings. We use all yields from 12 to 120 months that are available from the data set provided by Liu and Wu (2021). The sample ranges from January 1990 to December 2021. "ADF" stands for an Augmented Dickey-Fuller test. We report the p-values for each factor and two different versions of the test.

Statistic	β_1	β_2	β_3
Mean	6.466	-1.932	-1.257
Standard Deviation	2.912	1.865	2.598
Minimum	0.731	-5.579	-8.326
25% Percentile	4.595	-3.292	-2.570
50% Percentile	6.280	-2.017	-0.857
75% Percentile	8.235	-0.661	0.494
Maximum	14.023	5.677	6.274
ADF (constant only)	0.846	0.005	0.030
ADF (constant + linear time-trend)	0.091	0.029	0.022

Figure 4: Realized vs implied excess bond returns. The blue solid line shows the one-year excess bond returns measured from data, following (1). The red dashed line displays the returns that would have been observed if yields followed (4) and the realization of the factors were the ones we estimated, as in Figure 3.



1, a natural question is how well the excess bond returns implied by the dynamics of the Nelson-Siegel factors track the returns we observe from the realized zero-coupon yields.

Figure 4 answers this question. The blue solid line represents the one-year excess bond returns computed using the observed zero-coupon yields, as defined in (1). The panel on the left shows returns for $n = 2$ years, while the panel on the right shows returns for $n = 10$ years. The red dashed line shows the one-year excess bond returns *implied* by our factor estimates, i.e., we plug our factor estimates into the right-hand-side of (6). Hence, the dashed line represents the excess bond returns we would have observed if, in fact, the zero-coupon yields perfectly followed the estimated Nelson-Siegel model. We deem our Nelson-Siegel approach successful in fitting the yield curve since both lines are practically the same in both panels, which is a manifestation of the flexibility provided by the approach from Nelson and Siegel (1987) coupled with the parametrization from Diebold and Li (2006).

It is worth noting that we do not impose the no-arbitrage restrictions that appear in some affine term-structure models, like Ang and Piazzesi (2003). On the one hand, these restrictions might increase

the efficiency in the estimation of factors. On the other hand, it is unclear how useful they are in the context of forecasting. Our goal when using the model in (4) is to fit the yield curve as well as we can at a given point in time and then analyze forecasts of these factors, which are ultimately tied to the predictability of excess bond returns as shown by Proposition 1. It is not straightforward that extra restrictions would improve the fit. Additionally, Diebold and Li (2006) note that if the no-arbitrage condition is approximately verified in the data, a flexible model would also generate fitted monthly yield curves that approximately respect no-arbitrage restrictions.[21]

In Appendix C, we further discuss alternative estimation procedures for the Nelson-Siegel factors and compare them. We also show how our choice for λ is very close to the optimal constant λ throughout the sample. Additionally, we show how the proposed reduced-form model beats a benchmark polynomial model in our sample in terms of fitting errors.

5 Dissecting Violations of the Spanning Hypothesis

The model in (4), together with Proposition 1, offers a natural way to test whether there are asymmetries in the violation of the Spanning Hypothesis across maturities and where they are coming from. The shorter end of the yield curve is more heavily influenced by β_2, while the intermediate and long ends are more influenced by β_3 and β_1, respectively. Under the Spanning Hypothesis, and given that these factors are linear combinations of yields due to our estimation strategy, all information needed to forecast these β's should be spanned by the zero-coupon yields themselves. One way to test this property is asking whether macroeconomic data can enhance the forecast of these factors. The asymmetry shown in Figure 1, coupled with Proposition 1, suggests that macroeconomic data should enhance the forecast of the shorter end of the curve *more than* the longer end. In this section, we use different methodologies to forecast the Nelson-Siegel factors out-of-sample and evaluate how useful the macroeconomic signals can be.

5.1 Regressions with Principal Components

Our first analysis considers the following linear forecasting model:

$$\beta_{i,t+12} = \alpha_i + \theta_i' C_t + \gamma_i' PC_t + \epsilon_{i,t+12}, \quad i \in \{1,2,3\}. \tag{8}$$

We deploy the same expanding window forecasting exercise from Section 3, but now we directly target β's. We follow the tradition in the forecasting and Machine Learning literature and focus on the out-of-sample R^2 of our predictions:

[21] the evidence on how useful those restrictions are in a forecasting context is mixed, at best. Ang and Piazzesi (2003) and Almeida and Vicente (2008) argue that they are helpful, but the effects are small, while Duffee (2002) finds no gains. Carriero and Giacomini (2011) developed a formal test to analyze the usefulness of no-arbitrage restrictions. They find that the answer is largely dependent on the loss function adopted by the econometrician. For the case of a quadratic loss function as in our approach, they do not find large gains by imposing no-arbitrage restrictions.

$$R^2_{oos} = 1 - \frac{\sum_{t=t_0}^{T} \left(\beta_{i,t} - \widehat{\beta}_{i,t}\right)^2}{\sum_{t=t_0}^{T} \left(\beta_{i,t} - \overline{\beta}_{i,t}\right)^2}, \tag{9}$$

where $\widehat{\beta}_{i,t}$ is a particular forecast and $\overline{\beta}_{i,t}$ is a benchmark. Note that this measure is equal to one minus the MSE ratio, such that it can be negative if the forecast created by a given method is worse than the benchmark itself in terms of mean squared error.

A natural benchmark for us is the random walk since the factors are persistent time-series.[22] This is also an important benchmark because it is available under the Spanning Hypothesis. After all, the econometrician can always guess that the future yield curve will be the same as today's yield curve (and hence that the factors, which are linear combinations of yields, will be the same).

Additionally, since the factors are persistent series, we also target their innovations directly. This seeks to alleviate the concern that forecasting persistent time series with less persistent ones might generate poor performance for reasons unrelated to the Spanning Hypothesis. Instead of directly predicting their levels, we also perform the predictions of their one-year innovations:

$$\Delta\beta_{i,t+12} \equiv \beta_{i,t+12} - \beta_{i,t} = \alpha_i + \theta_i' C_t + \gamma_i' P C_t + \epsilon_{i,t+12}, \quad i \in \{1,2,3\}. \tag{10}$$

We then infer the predicted level by:

$$\widehat{\beta}_{i,t+12} = \beta_{i,t} + \widehat{\Delta\beta}_{i,t+12}. \tag{11}$$

As before, we have at least two natural choices for C_t: the forward rates or the lagged factors. We focus on results that control for the forward rates and leave most of the results controlling for lagged Nelson-Siegel factors for the appendix. The main conclusions do not rely on that choice. We prefer the results using the forward rates because their linear span is potentially richer than using the Nelson-Siegel factors, even though the 3-factor representation fits the U.S. yield curve remarkably well.

Table 3 displays results targeting the level of the factors in Panel A and their innovations in Panel B, in both cases controlling for the forward rates. In both panels we use the random walk as the benchmark. We also provide the p-values of a Diebold and Mariano (1995) test in which the null hypothesis is that the addition of macroeconomic information through the principal components does not enhance the forecasting of the factors (or their innovations). In both panels, the column labeled "No Macro" is our baseline specification when we impose $\gamma_i = 0$. Under the spanning hypothesis, as we further condition our forecasts on macroeconomic data, improvements in the forecasting of factors should only be due to sampling noise. We sequentially add principal components taken from the FRED-MD data set and keep track of the respective out-of-sample R^2 and associated p-values.

We start focusing on Panel A. The out-of-sample R^2 under the baseline is negative for all the three factors. As we increase the number of principal components taken from our macroeconomic data, we

[22] We have also estimated AR(1) models for the Nelson-Siegel factors and verified that a random walk is a harder benchmark to beat out-of-sample at the one-year horizon. Results are available upon request.

Table 3: We report the out-of-sample R^2 attained from the model in (8) for the baseline with no macroeconomic data included and with different numbers of principal components. Negative values imply we couldn't beat a random walk. We also show p-values to compare whether any improvement was statistically significant, comparing the "No Macro" baseline and the different forecasts. The first panel targets the level of the factors, while the second panel targets their innovations.

Panel A: Predicting Level													
		Number of Macro PCs						p-values					
Target	No Macro	1	2	3	4	5	8	1	2	3	4	5	8
β_1	-0.21	-0.17	-0.19	-0.15	-0.11	-0.09	0.03	0.18	0.33	0.13	0.11	0.10	0.01
β_2	-0.08	-0.08	0.17	0.22	0.21	0.23	0.22	0.49	0.01	0.02	0.02	0.02	0.05
β_3	-0.12	-0.15	-0.06	-0.07	-0.07	-0.07	-0.10	0.92	0.07	0.19	0.20	0.21	0.43
Panel B: Predicting Innovations													
$\Delta\beta_1$	-0.19	-0.15	-0.17	-0.14	-0.10	-0.08	0.05	0.19	0.32	0.17	0.12	0.10	0.01
$\Delta\beta_2$	-0.11	-0.12	0.14	0.18	0.17	0.19	0.18	0.52	0.00	0.02	0.02	0.02	0.05
$\Delta\beta_3$	-0.10	-0.12	-0.06	-0.05	-0.05	-0.06	-0.08	0.93	0.17	0.25	0.26	0.31	0.41

observe two different patterns. For both β_1 and β_3, the inclusion of macro data almost never makes the forecasting model better than a random walk out-of-sample, which was indeed a better model virtually across every possible scenario for β_1 and β_3. That is easy to see because almost all values for R^2 are negative. For β_2, in contrast, we observe a strong improvement in forecasting power, reaching an out-of-sample R^2 of more than 20% in some cases, which is statistically different than both our "No Macro" baseline and the random walk benchmark.

Panel B tells a similar story. The improvements in forecasting ability for β_2 when incorporating information from macro variables are much stronger than any improvements in β_1 or β_3. Interestingly, R^2 values attained are similar across both panels, showing that our results cannot be explained by the simple fact that yields are persistent.

The finding that macroeconomic data is helpful in forecasting β_2, but not the other factors, helps us understand the asymmetry documented in Figure 1. Since innovations in β_2 are disproportionately more important for bonds of shorter maturities, this pattern of predictability of the factors translates to stronger predictability of risk premium at the shorter end of the yield curve, in light of Proposition 1.

In Appendix B, Table B.5 shows the analogous results when controlling for the lagged Nelson-Siegel factors instead of the forward rates. The out-of-sample R^2 is almost never positive for either β_1 or β_3. In contrast, we get a positive R^2 of around 20% for β_2 across different specifications with macroeconomic data. Alternatively, Tables B.3 and B.4 show a similar pattern if we consider an in-sample exercise instead of an out-of-sample scheme. When we include principal components of the FRED-MD data set, the R^2 of the in-sample regressions increases relatively more for β_2 than for the other factors. The improvements for β_1, for example, are negligible.

Taken together, our evidence using principal components points to a pattern that is largely consistent with and sheds light on the asymmetry from Figure 1. The macroeconomic information spanned by the principal components of the FRED-MD data set is more helpful to forecast β_2 than the other factors, i.e., the violations of the Spanning Hypothesis are stronger at the shorter end of the yield curve.

5.2 Regularized Linear Models

Although simple to use and to communicate, principal component analysis has at least two drawbacks in our context. First, they fall into the "unsupervised" category of techniques for large data sets, as described by Hastie et al. (2009). This means that the dimensionality reduction they provide is not necessarily designed to improve forecasting. It might be the case that a certain linear combination of the different variables can explain a large amount of the total variation of the data but does not enhance the forecast for a given target. The choice of the target is irrelevant to the extraction of the principal components.

The second drawback is the lack of interpretability. By definition, principal components will use information from all variables in the data set. Ludvigson and Ng (2009) analyze how these principal components load on different variables and argue that some of them are related to real activity measures and inflation. However, that interpretation is only tentative and there is no reason for this result to hold over time or in different sub-samples. Moreover, our out-of-sample design requires sequential extraction of principal components. If we were to follow the same path, we would have to analyze the rotations implied by different principal components for each out-of-sample forecast we make, which is not feasible.

To avoid both drawbacks and further inspect the asymmetry in the violations of the Spanning Hypothesis, we stay in the realm of linear forecasting models but leverage regularization techniques. These methods are common in the Machine Learning literature and tend to be used in forecasting exercises when there is a large number of covariates. We focus on the Ridge (Hoerl and Kennard, 1970), Lasso (Tibshirani, 1996), and Elastic Net (Zou and Hastie, 2005) methods.[23] These methods have recently been used in forecasting exercises as in Gu et al. (2020), Medeiros et al. (2021), Bianchi et al. (2021) and Feng et al. (2022). They have also been coupled with standard inferential theory in the context of factor models for equity returns by Feng et al. (2020) and Giglio et al. (2021).

We let $X_t = [C_t'; F_t']'$ denote a vector containing variables that span the yield curve (C_t) and all the columns from the FRED-MD data set (F_t). We will still predict any target with a linear combination of variables in X_t. However, we will estimate this linear combination by minimizing a loss function that penalizes both in-sample forecasting errors and the "size" of the vector providing the optimal linear combination. Assuming we are targeting β_i, for given non-negative scalars $\psi_1, \psi_2 \geq 0$, we minimize:

$$\min_{\alpha_i, \gamma_i} \left\{ \frac{1}{T-12-t_0} \sum_{t=t_0}^{T-12} \left(\beta_{i,t+12} - \alpha_i - \gamma_i' X_t\right)^2 + \psi_1 ||\gamma_i||_1 + \psi_2 ||\gamma_i||_2 \right\}, \tag{12}$$

where $||.||_j$ denotes the L^j-norm of a vector for $j = 1, 2$ and time runs from t_0 to T in a generic sample. We then predict:

$$\widehat{\beta}_{i,t+12} = \widehat{\alpha}_i + \widehat{\gamma}_i' X_t. \tag{13}$$

We will also target the innovations, as we did before. In that case, analogously, we predict the innovations out of sample and define the forecast for the level as the lagged level plus the predicted

[23] See Hastie et al. (2009) for an in-depth treatment of these methods.

innovation, exactly as in (11).

This notation encompasses the three regularized models we consider:

1. $\psi_1 = 0, \psi_2 > 0 \implies$ Ridge
2. $\psi_1 > 0, \psi_2 = 0 \implies$ Lasso
3. $\psi_1, \psi_2 > 0 \implies$ Elastic Net

Even though these models might look similar, they behave differently. The Ridge model is the simplest of the three. The L^2-penalization will force coefficients of very correlated variables to be close to each other. It will not, however, make these coefficients be exactly zero. In that sense, Ridge is the only model that will not perform model selection from the three options above.[24] It will try to use all the information available, attaching similar weight to variables that are correlated and might span similar information. All estimates will be shrunk towards zero - or "regularized". The degree of shrinkage is controlled by the scalar ψ_2.

Lasso, on the other hand, will set several coefficients to exactly zero. That is due to the lack of smoothness implied by the L^1-norm. In general, it will work well in environments where a few signals from X_t can generate a good forecast for the given target, but they are hidden among several irrelevant ones. The penalty incurred by setting a given coefficient to a value different from zero is controlled by ψ_1. The greater this value, the more zeros $\widehat{\gamma}_i$ will contain - the more "sparse" $\widehat{\gamma}_i$ will be.

Finally, the Elastic Net is a joint estimation procedure that will impose *both* sparsity and shrinkage since the penalty function is a linear combination of norms. The price to pay for such flexibility is that two different hyperparameters need to be estimated. This is a non-trivial task since, *a priori*, there is no recommended number to which we can set these values. And, importantly, forecasting performance crucially depends on them. [25]

As we did in our out-of-sample exercise with principal components, we adopt an expanding window framework. Our first forecast was made for January 1990, and our last forecast was made for December 2021, so we computed 384 out-of-sample forecasts in total. For any point in time, we divide the available data into two consecutive parts: the estimation (or "training") sample and the validation set. The estimation sample is used to numerically solve the optimization in (12) for a given pair (ψ_1, ψ_2). With those estimates, we use (13) to forecast the observations contained in the validation set. We can then compute forecast errors and compute the mean squared error *within the validation set*. This measure is associated with the specific pair (ψ_1, ψ_2) then used. We minimize the mean squared error in the validation set by picking the best possible candidate combination (ψ_1, ψ_2) from a user-specified grid using a simple grid-search method. We let the validation set represent 20% of the data available at a given point

[24] We use the term "model selection" to denote the ability of a method to automatically pick a subset of variables out of a larger set of options and predict based on the chosen set.

[25] One can also adopt a Bayesian interpretation of these estimators, as Giannone et al. (2021) highlight. They numerically coincide with the mode of the posterior distribution of parameters if we assume that the targets are conditionally Gaussian and we set specific priors for the coefficients. For example, the Ridge method coincides with the case of a Gaussian prior on γ_i, while Lasso is equivalent to imposing a Laplacian prior in this setting. The Elastic Net corresponds to a prior that mixes both distributions. The tightness of these priors is controlled by the penalty parameters ψ_1 and ψ_2, respectively.

in time, using 80% for estimation. Since we adopt an expanding window, both the estimation sample and the validation set increase in size as we move ahead in time.[26]

5.2.1 Forecasting Ability

We compare the out-of-sample R^2 with and without the macroeconomic variables across the three different models and different targets, displayed in Table 4. In Panel A, we let C_t stack the forward rates, and in Panel B, we use the lagged Nelson-Siegel factors as a way to span the information from the yield curve. For each panel, the first three columns display the performance for each method when $X_t = C_t$, which is our baseline case. The second set of three columns shows the performance when we use all the available variables in the FRED-MD data set. The last set of columns reports the p-value of a Diebold and Mariano (1995) test in which the null hypothesis is that the forecasting ability is the same with and without macroeconomic data.

We start by analyzing predictions of innovations of the factors in Panel A. Across the three factors, we see that the forecasting ability is similar across the three methods. For β_1, the baseline R^2 is around 12%, which implies that the baseline model outperforms a random walk. However, the best one can do by allowing for macroeconomic data to enter the forecasting model is also 12%, such that there is no sign of improvement in out-of-sample forecasting. The situation is even worse for β_3, where the inclusion of macroeconomic variables decreases forecasting performance for all three methods.

The results for β_2 are in sharp contrast to those for β_1 and β_3. Under the "No Macro Data" baseline, we have essentially the same performance as a random walk. In contrast, the addition of macroeconomic variables brings the out-of-sample R^2 to around 20%, which is statistically higher than the baseline results at all usual confidence levels. This is the "supervised" version of the previous result using principal components: macro data can only improve the forecast for the short-run factor β_2.

The bottom three rows of Panel B repeat this exercise but control for the lagged Nelson-Siegel factors. Again, the spanning hypothesis is only violated through β_2. Another interesting pattern is that the Ridge method seems less efficient than the other two when faced with the high-dimensional panel of macroeconomic variables. Lasso and Elastic Net performed similarly but could beat Ridge, even though by a small margin. We believe this is intuitive since some of the FRED-MD predictors might be irrelevant for yield curve forecasting, and the L^1-norm penalization makes the associated coefficients exactly zero.[27]

Table 4 also shows a similar qualitative story when we focus on the results targeting the levels of factors. The only case in which we can beat a random walk is when we allow the methodology to use all macroeconomic variables and we target β_2. However, the absolute forecasting performance is worse than the cases in which we predict innovations. In particular, it is much worse for β_1. This is related to the fact that these factors are persistent, and regularized models are known to work empirically worse in

[26] This validation procedure is different from standard cross-validation typically employed in the Machine Learning literature (see Hastie et al., 2009). The temporal dimension of our setting makes us unable to use methods assuming observations to be independent. It is, however, very similar to the approach of Bianchi et al. (2021). See Arlot and Celisse (2010) for an in-depth discussion of different validation methods.

[27] Our (unreported) results show that on average, for both Lasso and Elastic Net, the typically chosen model here uses around 15-25 variables from the FRED-MD data set, completely ignoring the other ones.

Table 4: R^2_{OOS} from different models and for different targets, using the random walk. We target both the level of the factors and their innovations, which are then added to the lagged values to compute the implied forecast. The first three columns display results for our baseline case where we use no information from the macroeconomic variables. The out-of-sample period starts in January 1990 and ends in December 2021. We use the regularized models as in (12) to make the forecasts. The penalization constants ψ_1, ψ_2 are chosen using a validation set that contains 20% of the available data at each point in time. Panel A controls for the forward rates, while Panel B lets C_t store the three lagged Nelson-Siegel factors. The last columns report the p-value for a test (Diebold and Mariano (1995)) whose null is of equal predictive ability between the baseline models and the models with macroeconomic data.

Panel A: Conditioning on Forward Rates									
Target	No Macro Data			All Macro Data			p-value		
	Ridge	Lasso	Elastic Net	Ridge	Lasso	Elastic Net	Ridge	Lasso	Elastic Net
β_1	-4.84	-4.82	-4.69	-4.06	-4.30	-4.18	0.00	0.00	0.00
β_2	-0.08	-0.13	-0.19	0.07	0.07	0.06	0.05	0.00	0.01
β_3	-0.41	-0.59	-0.59	-0.47	-0.46	-0.45	0.78	0.04	0.03
$\Delta\beta_1$	0.12	0.12	0.09	0.01	0.12	0.12	0.96	0.50	0.27
$\Delta\beta_2$	0.01	-0.02	-0.01	0.15	0.22	0.19	0.02	0.00	0.00
$\Delta\beta_3$	0.04	-0.02	-0.03	-0.13	-0.09	-0.08	1.00	0.95	0.95

Panel B: Conditioning on Lagged Nelson-Siegel Factors									
Target	No Macro Data			All Macro Data			p-value		
	Ridge	Lasso	Elastic Net	Ridge	Lasso	Elastic Net	Ridge	Lasso	Elastic Net
β_1	-4.91	-4.73	-4.81	-3.76	-5.08	-4.53	0.00	0.97	0.10
β_2	0.00	-0.12	-0.12	0.08	0.07	0.02	0.16	0.00	0.08
β_3	-0.41	-0.47	-0.49	-0.45	-0.35	-0.39	0.71	0.04	0.09
$\Delta\beta_1$	0.12	-0.00	0.11	-0.29	0.04	0.08	1.00	0.30	0.84
$\Delta\beta_2$	0.10	0.08	0.12	0.18	0.25	0.24	0.11	0.00	0.01
$\Delta\beta_3$	0.08	0.04	0.02	0.00	0.03	0.07	0.95	0.70	0.02

such scenarios.[28] It is, nonetheless, reassuring that our main finding also arises in this more challenging case.

Overall, the evidence from the regularized models agrees with the evidence from our approach using principal components. No matter how we span the yield curve or how we target β_2, we detect violations of the Spanning Hypothesis. On the other hand, violations through β_1 and β_3 are much smaller and often inexistent. In general, forecasting the innovations, i.e., the first differences, proved to be better in terms of pure predictive performance. We also find that results were similar across the regularized models, with the Ridge method slightly underperforming Lasso and Elastic Net when macroeconomic data is added.

5.2.2 Model Selection

As mentioned above, Lasso and Elastic Net impose sparsity in $\widehat{\gamma}_i$, i.e., some (or most) coefficients will be set to zero. In that sense, the loss function we chose will effectively select a forecasting model.

[28] Figure A.6 in Appendix A further illustrates this point.

While the analysis using principal components was virtually silent regarding what variables were more important for forecasting, the regularized models are more explicit on that front. Since we have to solve (12) numerically each time we perform an out-of-sample forecast, we can keep track of the choices made by these methods. Then, we can use the official classification of these variables from the St. Louis Fed to aggregate this information at the group level.[29]

We compute how frequently variables from each group were chosen and show results in Figure 5 for the three Nelson-Siegel factors. We count how many choices were made in total over time and what percentage of those choices can be attributed to each group. For example, if a given model were to pick only labor market indicators and price measures in equal proportion, the pie charts would show these two groups with 50%-sized slices. The pie charts in the first row show results for Lasso and the ones in the second row for Elastic Net. For β_1, we show results for the innovations due to the poor performance when targeting its level, as shown in Figure A.6. For the other two factors, we focus on choices made when we target the levels. We focus only on results when we control by forward rates.[30]

A first evident pattern is that indicators related to price levels were the most chosen variables for β_1 both by the Lasso and by the Elastic Net. These are different inflation measures, consumption expenditure indexes, and commodity prices. We see these variables as all linked to the current state of inflation. This result seems intuitive to us. Higher inflation typically means that the monetary authority will have to increase the short-term interest rates. Since inflation is relatively persistent and monetary policy acts with a lag, agents might infer that short-term interest rates will have to remain higher for some time, impacting also the longer-term interest rates as well.

For β_2, on the contrary, we find that no group is particularly dominant, with almost identical results for Lasso and Elastic Net. This suggests that the methodology is mixing signals from different groups. We highlight, however, that price indexes, labor market indicators, and housing construction indicators represent more than 50% of all choices using Lasso and almost 60% of all choices for Elastic Net.[31] This means that both methodologies are extracting information from signals that are typically informative about general business-cycle conditions. Such information is strong enough to make these forecasts more precise than both the baseline case under the Spanning Hypothesis and a random walk. The shorter end of the yield curve, summarized by the short-run factor β_2, is more likely to be directly affected by monetary policy decisions. Since these decisions are taken conditionally on the business cycle, it seems natural that a wide array of signals that are informative about the current state of the economy helps forecast this factor. A similar pattern arises for β_3, for which both methodologies end up mixing the different signals available in the data set, even though this is not enough to significantly improve its forecast.

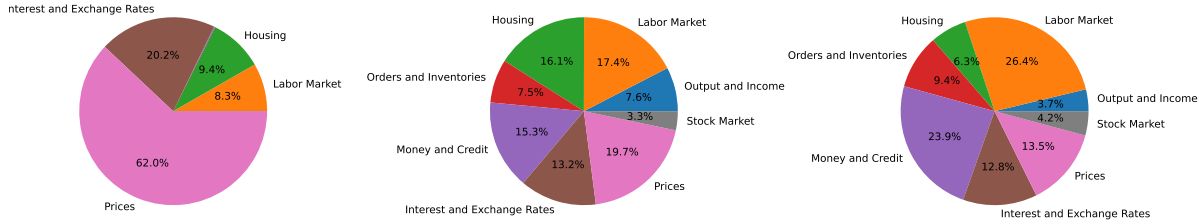
[29] See our Appendix D for a full list of variables and their classification.

[30] Controlling by the Nelson-Siegel factors leads to very similar pie charts, which we omit for the sake of space but are available upon request.

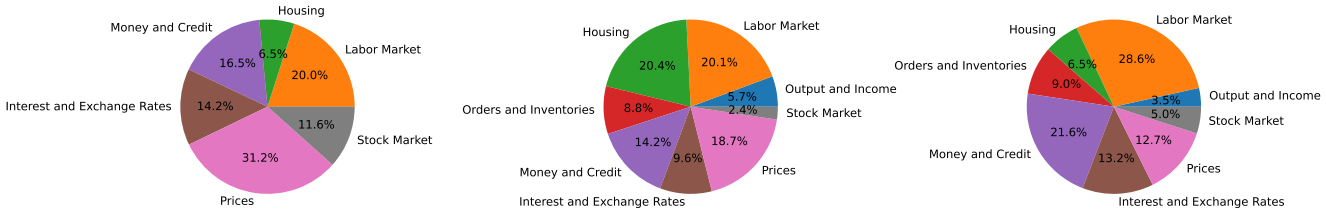
[31] The average number of chosen variables from the FRED-MD dataset is 10-15 across both methodologies. Giannone et al. (2021) analyze how these regularization methods perform when making forecasts in different contexts in Economics and finds that the best forecasts are generally made by mixing signals and not picking just a few ones. This is a phenomenon they call the "illusion of sparsity". In our view, the evidence that the forecast for β_2 is mixing information from the different groups is a manifestation of this lack of sparsity.

Figure 5: Most frequently chosen groups. For each group of variables, we keep track of how many times variables of that group were chosen either by Lasso or by the Elastic Net. Then, we compute this number as a fraction of the total choices. The top row shows results for the Lasso, while the bottom row shows results for the Elastic Net. We let C_t store the lagged value of the forward rates. The out-of-sample period is January 1990 to December 2021.

(a) Lasso $(\beta_1, \beta_2, \beta_3)$



(b) Elastic Net $(\beta_1, \beta_2, \beta_3)$



5.3 Allowing for Non-linearities

The different forecasting methods used so far are fundamentally linear. The empirical evidence from Gu et al. (2020), Medeiros et al. (2021), and Bianchi et al. (2021), for example, highlights that non-linear methods can, in the context of financial and economic forecasting, significantly improve on linear methods. Although our main concern is *not* on what type of methodology is the best for pure forecasting, we also want to rule out the possibility that the results so far are due to the linear nature of the methodologies we used. With that goal in mind, we mainly focus on the Random Forest methodology to allow for non-linearities in forecasting. Aside from being computationally simpler than a Neural Network, Medeiros et al. (2021) find that it performs the best when forecasting inflation (which is also a persistent target, as the Nelson-Siegel factors). In a similar fashion, Goulet-Coulombe (2023) does an in-depth

analysis of their use (and success) when predicting macroeconomic variables.[32] We focus not only on the forecasts generated by this methodology but also on the *feature importance* of different variables, which is a measure of how useful each of the signals is.

5.3.1 Forecasts

To implement this model, we adopt the standard CART algorithm from Breiman (2001).[33] A random forest is composed of several regression trees, in our case. A tree is a collection of split points chosen to minimize the in-sample mean-squared error in prediction. The CART algorithm is *greedy* in the sense that it searches for the best variable and split point at each step that will minimize the prediction error up to that stage. That leads to regression trees that are estimators of the conditional mean of a given targeted variable that have low bias but high variance. We can construct several different trees at each prediction step and average their individual predictions. The trees differ because the subset of variables used in their construction is random, although of fixed size. This implies that the predictions of different trees typically display low correlation, which is useful since the final averaging delivers an estimator with lower overall variance.[34] At each time t, we reestimate the forest and predict both factors and their innovations one year ahead.

Table 5: We report the out-of-sample R^2 from the Random Forest technique using the random walk as the benchmark. The out-of-sample period is 1990-2021. We used 500 trees at each point in time. We target both factor levels and their innovations. The p-values assess whether statistical improvements over the "No Macro" baseline are significant. Implementation followed Breiman (2001).

	Lagged Factors			Forward Rates		
Target	No Macro	All Macro	p-value	No Macro	All Macro	p-value
β_1	-1.48	-1.93	0.87	-0.76	-0.72	0.39
β_2	-0.08	0.27	0.01	-0.34	0.23	0.00
β_3	-0.41	-0.16	0.02	-0.58	-0.22	0.01
$\Delta\beta_1$	-0.17	0.00	0.05	-0.53	-0.04	0.00
$\Delta\beta_2$	-0.08	0.32	0.00	-0.42	0.32	0.00
$\Delta\beta_3$	-0.37	-0.01	0.02	-0.33	-0.25	0.25

Table 5 displays the out-of-sample R^2 with and without macroeconomic variables, with the respective p-values. The columns from the left use the lagged Nelson-Siegel factors as controls, and the columns from the right use the forward rates as controls. The pattern is clear no matter how we control for the information already in the yield curve: adding macroeconomic variables only significantly improves forecasting power for the short-run factor β_2. In contrast, we can never beat a simple random walk for β_1 and β_3.

Across the board, we find that predicting the innovations provides better forecasting performance.

[32] We refer the reader to Bianchi et al. (2021) for a detailed study on the optimal design of neural networks to predict bond risk premia, which we do not tackle here.

[33] We use the standard implementation available in `scikit-learn` in Python.

[34] For a full treatment of the Random Forest, see Hastie et al. (2009). At each time t, we build 500 trees to construct our forest. The number of variables that are selected for each tree is always a third of the total amount. This is the recommended choice by Hastie et al. (2009) for the case of regression trees.

In fact, across all our estimation strategies so far, coupling macroeconomic data and targeting $\Delta\beta_2$ was the only strategy that delivered an R^2 greater than 30%. This echoes the results in Medeiros et al. (2021) regarding the efficiency of Random Forests in using the small amount of data provided to it while still allowing for non-linearities.

5.3.2 Feature Importance

What is the relative importance of information from the yield curve versus information from the macroeconomic variables in the Random Forest? When building an individual regression tree, the CART algorithm readily delivers a statistic called *feature importance* for each of the variables used in the prediction exercise. This is a measure of how important each of the available variables was for the reduction of the in-sample MSE, starting from a baseline value that just guesses the sample mean of the targeted variable.

If we denote by $f_{i,j,t}$ the feature importance for variable i when building tree j at time t, we can define the importance of variable i, denoted by F_i, as the average across trees and time:

$$F_i \equiv \frac{1}{384} \cdot \sum_{t=\text{Jan, 1990}}^{\text{Dec, 2021}} \frac{1}{N_i} \cdot \sum_{j=1}^{N_i} f_{i,j,t},$$

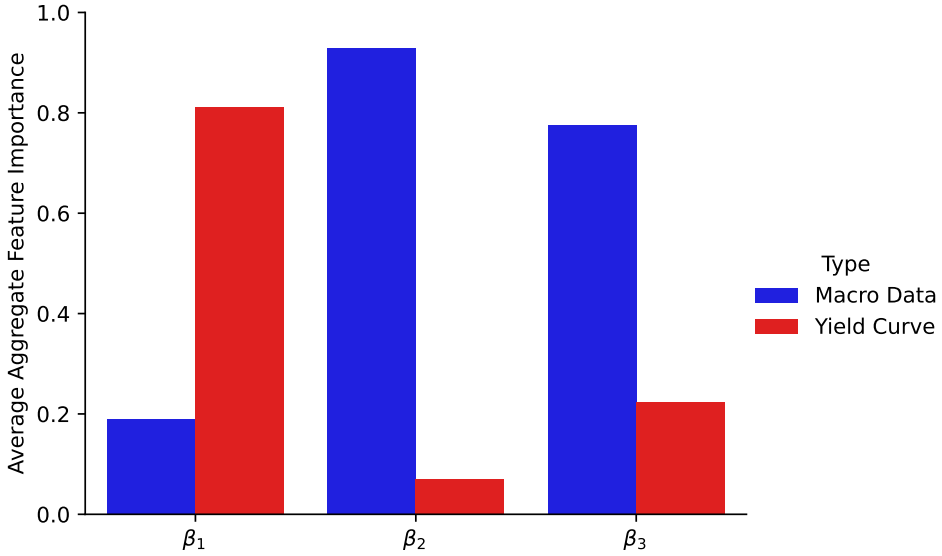
where N_i is the total number of trees in which variable i was used. Given a collection of feature importances F_i, we can normalize everything so $\Sigma_i F_i = 1$. Hence, the normalized feature importance represents, on average, how much each variable contributed to the in-sample MSE reduction, i.e., how helpful each variable was in predicting the Nelson-Siegel factors.

We classify the variables into two types: the ones that come directly from the yield curve data and the ones that come from the FRED-MD data set. Figure 6 displays, for each factor, the aggregate feature importance. We focus on results that control for the forward rates and target the factors directly, but the same conclusion holds for other specifications. For each factor, the blue and red bars sum up to one.

For β_1, roughly 80% of the in-sample MSE is attributed to variables that come from the yield curve (the ten forward rates). On the other hand, the behavior for β_2 is exactly the opposite. More than 90% of the MSE reduction is attributed to the macroeconomic variables, as indicated by the dominance of the blue bar over the red one. This is at odds with the Spanning Hypothesis. The medium-run factor β_3 displays an intermediate behavior. However, as seen from Table 5, any improvement brought by the addition of macro data to the forecasting model was not enough to beat a random walk, which is available under the Spanning H ypothesis.

Figure A.7 in Appendix A displays the average feature importance F_i for each variable and each factor, keeping the same color coding. In line with the aggregation reported in Figure 6, the red bars dominate the plot for β_1, while they are very small for β_2. Additionally, one can see that there is no single variable from the macroeconomic data set that brings most of the forecasting improvement alone. The methodology picks a mix across signals that are individually weak, but that generate sizeable forecasting gains when combined.

Figure 6: The figure displays the average normalized feature importance of each group of variables when predicting the Nelson-Siegel factors. The red bars aggregate information from the yield curve (forward rates here). The blue bars aggregate measures of feature importance from the FRED-MD dataset. The out-of-sample period is 1990-2021.



The results for the Random Forest are, again, in line with our previous evidence. Adding macroeconomic variables to the predictive model only significantly improves the forecasting power for β_2. The information contained in the yield curve is, in relative terms, much less important for the forecast of the short-run factor than for the long-run factor. We see this as another manifestation of the asymmetry in violations of the Spanning Hypothesis.

6 Expected Short Rates vs Term Premium

Zero-coupon yields can always be decomposed into two components: the expected path of short rates and the term premium. Such decomposition is an accounting identity (Gürkaynak and Wright, 2012) that comes from the definition of excess bond returns and holds regardless of the underlying data-generating process. As we prove in Appendix E, using the definition in (1) we can write:

$$y_t^{(n)} = \underbrace{\mathbb{E}_t\left[\frac{1}{n}\cdot\sum_{k=1}^{n} y_{t+12\cdot(k-1)}^{(1)}\right]}_{\text{Expected average short rate } \left(ES_t^{(n)}\right)} + \underbrace{\mathbb{E}_t\left[\frac{1}{n}\cdot\sum_{k=1}^{n} xr_{t+12\cdot k}^{(n-k+1)}\right]}_{\text{Term premium } \left(TP_t^{(n)}\right)} \tag{14}$$

The first term reflects the expectation of the average yield an investor would earn through a sequence of n 1-year risk-free investments. The second term represents a wedge between this expectation and the yield, at time t, earned by an n-year bond, which is $y_t^{(n)}$. Such a wedge is compensation for the risk an investor takes by choosing the longer investment option instead of a series of 1-year investments over time.

Our empirical exercises focused on using the information available up to time t to predict either returns or linear combinations of yields (the Nelson-Sigel factors). Equation (14) holds period by period

and, therefore, we can also write that $y_{t+12}^{(n)} = ES_{t+12}^{(n)} + TP_{t+12}^{(n)}$. It is natural to ask what component of yields at $t+12$ we are predicting better when using additional information from the macroeconomic signals up to time t. This exercise is also informative about the nature of the violations of the Spanning Hypothesis. The standard interpretation of these components is that ES_t is more heavily influenced by monetary policy, for example, while TP_t depends on factors such as the current level of risk-aversion displayed by market participants and their ability to hold long-duration bonds. For instance, if all investors were risk-neutral, the second term in (14) would be identically zero across all maturities.

Even though this decomposition holds very generally, actually computing these expectations requires a model. Dynamic Term Structure Models are largely built to discipline this decomposition. Here, we adopt the model from Adrian et al. (2013) to acquire, for each maturity, separate time series for ES_t and TP_t[35]. Figure A.8 in Appendix A displays such time series for $n = 2$ and $n = 10$. One important characteristic of this model is that both the ES_t and TP_t measures are generated in an environment featuring the Spanning Hypothesis. Then, macroeconomic signals should not enhance the prediction of either of those components after we condition the current yield curve.

We use the different methodologies we have presented so far and try to forecast both time series one year ahead, first conditioning only on the forward rates at time t and then on the forward rates aided by macroeconomic signals. We do this separately for $n = 1, 2, ..., 10$, using the same out-of-sample period (1990-2021). We track the mean-squared error in prediction with and without macroeconomic variables and plot the ratio between these quantities for different maturities in Figure 7. When this ratio is below 1, it implies that macroeconomic data was helpful in improving forecasts; when it is above 1, it implies that the baseline forecast without any help from macroeconomic signals was more accurate.

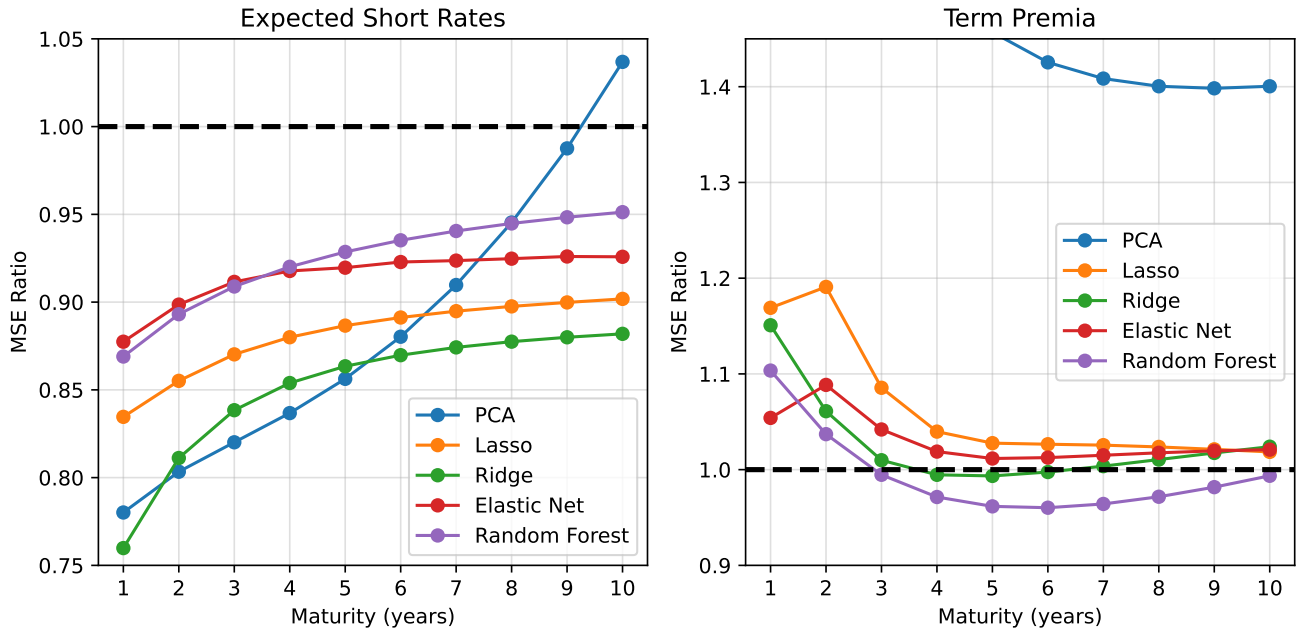
The left panel presents results for the ES_t component, while the panel on the right shows results for TP_t. There is a sharp contrast between the two panels. First, across different methodologies, the macroeconomic signals improved the forecasts of the expected path of short rates (ES_t), which was not the case for TP_t. Using an extended information set hurt the prediction of TP_t. Second, there is also an asymmetry in predictability in the left panel. The improvements due to the macroeconomic signals are stronger for shorter maturities. For instance, the Ridge methodology delivers almost a 20% reduction in MSE for $ES_t^{(2)}$ but just about 10% for $ES_t^{(10)}$.

The empirical result displayed in Figure 7 is unequivocal: the forecast precision improvements delivered by the macroeconomic variables happen because they improve the prediction of ES_t. The different methodologies extract information from the FRED-MD dataset that is useful to predict the path of the short rates in the U.S., and not the TP_t component.

This pattern also helps us interpret the evidence from Figures 1 and 2. The ES_t term represents a larger fraction of the observed yields for the shorter maturities, with a minimal contribution of TP_t. As the maturity increases, TP_t becomes larger as a fraction of the observed yields. This can be seen in Figure

[35]This model is the same one used by the New York Fed to generate their "official" measure of the term premium. In contrast to other competing models of the literature, it is easier to estimate and can handle noisy measurements of yields. Since it is not estimated by maximum likelihood methods, it also allows for more general dynamics of underlying factors. It falls within the class of "essentially affine" models from Duffee (2002). We downloaded the fitted time series directly from the New York website.

Figure 7: The panels show the ratio between the mean-squared error in prediction with and without macroeconomic signals for each maturity when targeting ES_t (left) and TP_t (right). Different colors refer to different forecasting methodologies. The out-of-sample period is 1990-2021. "PCA" refers to a forecast done with six principal components from the macroeconomic variables. Penalty hyperparameters for Lasso, Ridge, and Elastic Net are chosen by cross-validation.



A.8 in Appendix A: the gap between $y_t^{(2)}$ and $ES_t^{(2)}$ is small, while it is larger for $n = 10$. The path of expected short rates is quantitatively more important for the shorter maturities[36]. Since that is precisely the part macroeconomic data can help with, the gains from using macroeconomic variables in prediction concentrate on shorter maturities. This explains the asymmetry in the prediction of bond returns and is consistent with violations of the Spanning Hypothesis through β_2.

7 Taylor Rule Deviations

So far, we have established that the violations of the Spanning Hypothesis happen at shorter maturities since they stem from the predictability of the short-run factor β_2, and they are due to improved forecasts for the path of short rates. Understanding why the macroeconomic signals are improving the forecast for the path of short rates is key to understanding the violations we document.

The first term in (14) is heavily influenced by monetary policy. For example, communication from the Federal Reserve informs market participants what the likely path for the Fed Funds rate is, which exerts influence on rates at shorter maturities more generally. And the content of such communication, as well as the actual decisions, is determined by business-cycle conditions. The macroeconomic signals we entertain are informative exactly about the current economic environment. From a purely statisti-

[36]Figure A.9 in Appendix A shows the average value of $ES_t^{(n)}/y_t^{(n)}$ over time for different maturities. As n increases, this ratio decreases on average, which is consistent with the idea that ES_t is quantitatively more important for shorter maturities. For $n = 1$, the average value is slightly above 1 due to the COVID pandemic, which is in our sample. In 2020, the term premium became negative. If we had stopped the sample in 2019, for instance, we would see all bars below 1.

cal point of view, it seems natural that conditioning on these macroeconomic variables improves the forecasts of expected short rates since they carry important information the Federal Reserve will use to decide the path of monetary policy in the future.

Although intuitive, such a channel does not exist in typical DTSMs. If we let X_t denote a vector of underlying risk factors, and i_t denote the risk-free rate for the shortest maturity in a model, one common assumption in DTSMs is that $i_t = \delta_0 + \delta_1' X_t$ for some coefficients δ_0, δ_1. Such mapping from risk factors to short rates has two important features in the context of a typical DTSM: 1) it is *known* to the agent whose stochastic discount factor prices bonds; 2) it is constant over time and does not depend on X_t itself. This first feature implies that the agent who is pricing the bonds has no uncertainty about monetary policy once the state X_t is known. Any uncertainty about future short rates exists only due to the uncertainty about future realizations of X_t. The second feature implies that the policy rate will be set using the same rule over time, leaving no room for changes in the assessment of the current economic conditions (e.g., transitory vs. persistent inflation).

In reality, however, the reaction function of the Federal Reserve lacks both of those features. It is unknown even to professional forecasters, and it changes over time. For example, Bauer et al. (2024) use survey data to identify a perceived monetary policy over time and characterize its substantial time variation. Jia et al. (2023) show that a large portion of the disagreement among forecasters is due to different perceptions about the central bank's Taylor Rule.

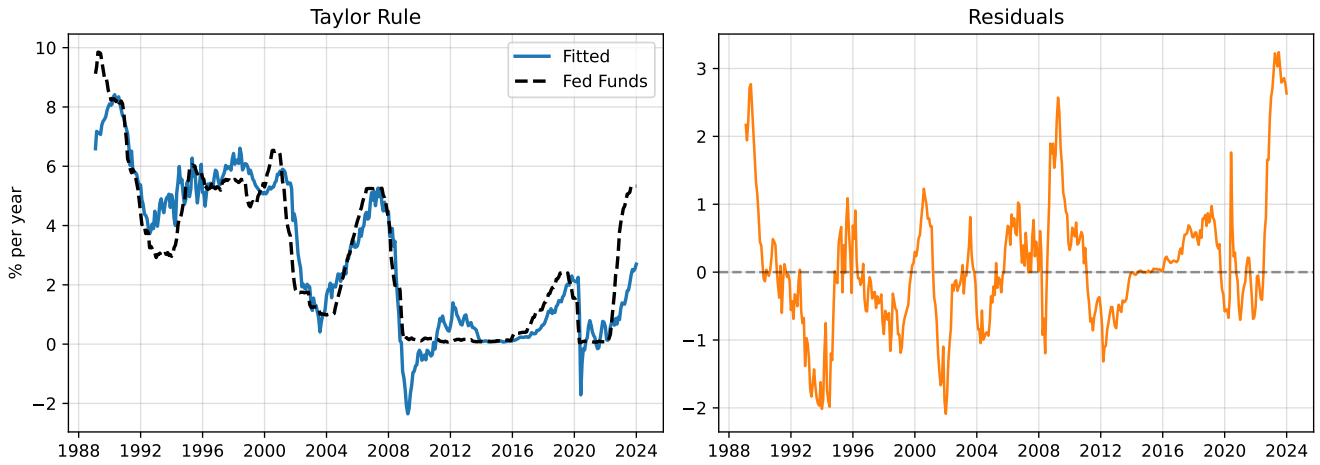
Additionally, even given a fixed monetary policy rule, there are different reasons why the policymaker could choose to deviate from it. They might be surprised by large shocks both domestically and internationally, or they could be prompted to act in response to developments in financial markets (Cieslak and Vissing-Jorgensen, 2021), for example. In any case, we interpret deviations from a certain monetary policy rule as an indication of some new aspect of monetary policy that was not present before.[37]

Recent evidence by Schmeling et al. (2022) validates our interpretation. They show that both forecast errors by professional forecasters and returns from Fed Funds futures positions are correlated to deviations from a Taylor Rule. When there is some new element guiding monetary policy, either due to a change in the assessment of current economic conditions or due to objectives not directly related to inflation and unemployment, even professional forecasters have to make sense of what the trajectory of short rates will be. Given some novelty in how monetary policy is conducted, forecasting becomes harder and errors become larger.

Our forecasting results from the previous sections motivate the following conjecture: *are macroeconomic variables especially useful to predict the short-run Nelson-Siegel factor β_2 when monetary policy deviates from "business as usual", i.e., when deviations from a Taylor Rule increase*? We now seek to test this conjecture. The first step is to actually fit a Taylor Rule for the Fed Funds rate and track deviations from it.

[37] We, by no means, aim at interpreting a deviation from a certain Taylor Rule as a monetary policy *shock*. The policy maker *chooses* when to deviate from a certain rule and has the freedom to conduct policy by any rule they want. In that sense, we do not take a stance on whether such deviations will or will not be anticipated by market participants. We only go as far as interpreting a deviation from a Taylor Rule as an expression of novelty in monetary policy given past decisions.

Figure 8: Left: the dashed lines represent the Fed Funds rate over time while the solid one represents our fitted rolling-window Taylor Rule, following (15). We use a 60-month rolling window. Right: the gap between the two lines denoted by ϕ_t.



7.1 Fitting a Taylor Rule

Given that the Federal Reserve's mandate is about inflation and unemployment, we fit a simple Taylor Rule using these two variables and use a 60-month rolling window to update the coefficients of the following regression:

$$i_t = \delta_0 + \delta_\pi \cdot \pi_t + \delta_u \cdot u_t + \nu_t \tag{15}$$

where i_t is the Fed Funds rate, π_t denotes quarterly core PCE inflation and u_t denotes unemployment. We start the sample in August 1979, when Paul Volcker became the chairman of the Federal Reserve. We define $\phi_t \equiv i_t - \widehat{i}_t$, where $\widehat{i}_t$ is the fitted value of such regression, as a deviation from the Taylor Rule. Hence, ϕ_t is computed with monthly data from the $[t-60, t]$ range. We follow Carvalho et al. (2021) and Schmeling et al. (2022) and estimate (15) by OLS.

The interpretation of ϕ_t is how much the policy rates deviate from what an observer could have expected given the last 60 months of data, assuming the Federal Reserve only takes inflation and unemployment into consideration when deciding the policy rate. Using rolling windows helps us accommodate coefficients that evolve over time following different crises and changes in the chairmanship of the monetary authority.

Figure 8 displays, on the left panel, the evolution of the Fed Funds rate as a dashed line and our fitted values as a solid line. The panel on the right displays the difference between the two curves, which is our measure ϕ_t of deviations. Figure A.10 in Appendix A shows this measure is not heavily influenced by our choice of using core PCE inflation, for instance. Using headline PCE would lead to a similar result. Extending the Taylor Rule to allow for quarterly and monthly inflation would also lead to similar results.

7.2 Can deviations predict higher precision?

With ϕ_t at hand, we can now test our conjecture. We define the loss-function when predicting the Nelson-Siegel factor β_i at time t as $L_{i,t} \equiv \left(\widehat{\beta}_{i,t} - \beta_{i,t}\right)^2$, where $\widehat{\beta}_{i,t}$ is a forecast done with or without macroeconomic data. For this exercise, we focus no the Random Forest methodology since it delivered the best results when predicting β_2 (see Table 5).

In order to match our rolling-window scheme to fit the Taylor Rule, we also use rolling-window Random Forest forecasts. Table 6 reports the out-of-sample R^2 when predicting the Nelson-Siegel factors with this forecasting strategy using a 180-month window. The results from this table show that our earlier evidence was not due to an expanding window forecast design. On average, even if we use rolling windows, violations of the Spanning Hypothesis happen only through β_2. Table B.8 in Appendix B reports the same exercise but using a 120-month window as a robustness check. The same result remains true.

Table 6: This table reports out-of-sample R^2 using the Random Forest method with 500 trees estimated at each point in time. The out-of-sample period is 1990-2021. Negative values mean we couldn't beat a random walk. We use 180 months for the rolling window. The p-values assess whether any improvement over the "No Macro" baseline is significant. The three first columns control for the information in the yield curve using lagged Nelson-Siegel factors, while the last three columns use forward rates as controls.

	Lagged Factors			Forward Rates		
Target	No Macro	All Macro	p-value	No Macro	All Macro	p-value
β_1	-1.20	-1.32	0.72	-0.63	-0.98	0.98
β_2	-0.07	0.20	0.02	-0.30	0.19	0.00
β_3	-0.47	-0.24	0.04	-0.67	-0.23	0.00

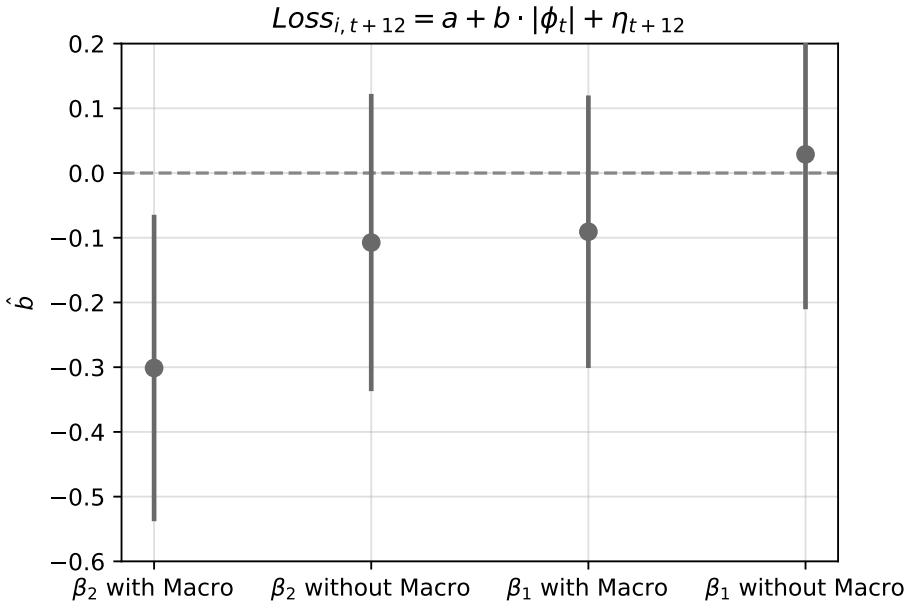
We conjecture that macroeconomic signals might be informative about monetary policy, and such information leads to better predictability at the shorter end of the yield curve. Arguably, such information might be more useful when there is some novelty in monetary policy that needs to be understood and processed by market participants. To test our conjecture, we run the following regression:

$$L_{i,t+12} = a + b \cdot |\phi_t| + v_{t+12}. \tag{16}$$

If $b < 0$, there is evidence that higher deviations from a Taylor Rule are associated with a lower value of the loss-function $L_{i,t}$, which is equivalent to higher precision in forecasting. If $b = 0$, such a correlation does not exist. Before running the regression above, we scale the loss function so b can be measured in standard deviations of the loss function per 1 p.p. (100 basis points) of absolute deviation.

Figure 9 plots the point estimates of b and associated 95% confidence intervals. We estimate the regression above for the forecasts of both β_1 and β_2, with and without macroeconomic data, to contrast results. The first estimate implies that 100 basis points worth of deviations from the Taylor Rule are associated with a *reduction* of 0.3 standard deviations of the loss function, i.e., an increase in precision. This implies that, when predicting the short-run factor, estimates done conditioning on macroeconomic information end up being more precise when deviations from the Taylor Rule are larger. The second

Figure 9: Point estimates for b from (16) with 95% confidence intervals using the HAC-estimator from Newey and West (1987). In each case, we scale $L_{i,t}$ so that b is measured in standard deviation per 100 basis points worth of deviations. The forecasts come from a rolling-window forecasting scheme that uses the Random Forest methodology.



estimate displayed in Figure 9 shows that the same phenomenon does not happen when we predict the short-run factor using only information from the yield curve. The estimated coefficient is around a third of the previous estimate, and we can't statistically reject it is zero. In a similar fashion, when we study the longer end of the yield curve, which is heavily influenced by β_1, we find no statistically significant association between precision and deviations from the Taylor Rule.

We take this evidence as favorable with respect to our conjecture. Indeed, forecasts computed using information from business-cycle variables from the FRED-MD dataset are more precise when there is some novelty in monetary policy, for which $|\phi_t|$ is a proxy. This effect is not present when we predict the longer end of the yield curve since it is generally less affected by monetary policy than the shorter end.

Figure A.11 in Appendix A displays the same pattern as Figure 9 but using the different versions of the Taylor Rule we considered (columns) and different controls for the information in the yield curve (rows). Across all the specifications, both the qualitative and quantitative patterns remain: there is a statistically significant reduction in the loss function (increase in precision) as the Taylor Rule deviations increase, but such effect is only present when predicting β_2 with macroeconomic information.

8 Discussion and Extensions

8.1 Previous Literature

Our results on the lack of predictability of term premia also speak to the previous literature. Ludvigson and Ng (2009) use principal components of the FRED-MD dataset to show that they can predict bond returns and push the idea that predictability comes through the term premium component. There

are several differences in our setup that might explain such a discrepancy. First, our decomposition of yields into ES_t and TP_t is disciplined by the model from Adrian et al. (2013), while Ludvigson and Ng (2009) fit a VAR to yields to recover ES_t. Second, our data covers more maturities and we use a longer sample. And third, we focus on (pseudo-) out-of-sample predictability.

We also speak directly to the literature on unspanned macroeconomic risk. Joslin et al. (2014) propose a model in which macroeconomic information can predict bond returns but is not revealed by yields. Such property arises because macroeconomic information affects expected short rates and term premia in an offsetting manner across maturities. This type of condition has been tested and rejected by the data through likelihood-ratio tests (Bauer and Rudebusch, 2017). Here, we show that macroeconomic signals from the FRED-MD dataset help us predict the path of short rates and not term premia. Hence, we find no evidence of this offsetting effect and see our result as complementary to Bauer and Rudebusch (2017).

8.2 More Non-linearity

Another way of allowing for non-linearities in forecasting is following a strategy similar to Ludvigson and Ng (2009) and considering higher powers of the principal components from the FRED-MD dataset. For example, for a given number K of principal components, we can define $PC_t^{(r)}$ as a $K \cdot r \times 1$ stacked vector of principal components and their respective powers up to the maximal exponent r:

$$PC_t^{(r)} \equiv \left[PC_{1,t} \quad \cdots \quad PC_{K,t} \quad PC_{1,t}^2 \quad \cdots \quad PC_{K,t}^2 \quad \cdots \quad PC_{1,t}^r \quad \cdots \quad PC_{K,t}^r \right]'$$

We can then use such variables in an out-of-sample predictive regression:

$$\beta_{i,t+12} = \alpha_i + \theta_i' C_t + \gamma_i' PC_t^{(r)} + \epsilon_{i,t+12}, \quad i \in \{1,2,3\}. \tag{17}$$

We report results for $r = 2$ on Table B.6 and for $r = 3$ on Table B.7 in Appendix B. The same qualitative behavior remains: there is extra predictability only through β_2.

8.3 Conditional Predictive Ability

In Appendix F, we adapt the framework from Giacomini and White (2006) to our setup and conduct a formal test of conditional predictive ability. Throughout the main text, whenever comparing forecasts with and without macroeconomic data, we focused on the *unconditional* null that forecasts done with the aid of macroeconomic signals could not be better, on average, than baseline forecasts. Another interesting question is whether an econometrician, with information up to time t, can decide whether to use or not use macroeconomic information to make the forecast. We show that forecasts using macroeconomic information end up being more precise when inflation is higher than its historical average. These are arguably times in which the Federal Reserve is more likely to act to tame inflation.

8.4 Other Explanations

Even though we have emphasized the role of macroeconomic information in helping the econometrician to get more precise forecasts for the path of short rates, we do not rule out other explanations for violations of the Spanning Hypothesis. For example, Cieslak (2018) finds that professional forecasters systematically fail to incorporate information about the business cycle in their forecasts. One possibility is that the information set of the econometrician is different from the information set of the representative agent.

Another possible explanation relies on alternative belief formation mechanisms by market participants. Piazzesi et al. (2015) show evidence that forecasters take both the level and the slope of the yield curve as more persistent than they actually are, which is consistent with extrapolative expectations. Hence, they might not extract the full extent of information present in the yield curve, leading to additional predictability by macroeconomic data.

In any case, we believe that stepping outside the full-information rational expectation framework is necessary to understand phenomena such as violations of the Spanning Hypothesis. We stress that the connection between DTSMs and uncertainty regarding the central bank's reaction function is relatively unexplored. In reality, agents need to learn this reaction function, and such learning seems relatively more important for the short end of the yield than for longer maturities.

9 Conclusion

We document asymmetric violations of the Spanning Hypothesis across bond maturities: macroeconomic variables contain unspanned predictive information about the shorter end of the yield curve. Using a Nelson-Siegel representation for yields and an implied decomposition of bond risk premia, we show that this pattern arises because macro data is useful for predicting a short-run factor related to the slope of the yield curve but not for predicting its level or curvature.

We introduce information from macroeconomic variables to our forecasting designs using regressions with principal components, regularization methods crafted to handle large-dimensional data sets, and fully non-linear methods like the Random Forest. All methodologies lead to the same qualitative conclusion.

The extra predictability of macroeconomic data brings steam from improved forecasts of the path of short rates and not term premia. Consistent with this idea, we show that the precision of forecasts made with macroeconomic data for the short-run Nelson-Siegel factor increases when the Federal Reserve deviates from a Taylor Rule. Such effects are ruled out by design from typical Dynamic Term Structure Models because a constant and known reaction function is typically assumed.

We believe that integrating uncertainty about monetary policy into Dynamic Term Structure Models is a promising area of research for both theorists and applied researchers. We leave this for future work, hoping it can help us resolve puzzles such as the failures of the Scanning Hypothesis.

References

Adrian, T., Crump, R. K., and Moench, E. (2013). Pricing the term structure with linear regressions. *Journal of Financial Economics*, 110(1):110–138.

Almeida, C. and Vicente, J. V. (2008). The role of no-arbitrage on forecasting: Lessons from a parametric term structure model. *Journal of Banking & Finance*, 32(12):2695–2705.

Altavilla, C., Giacomini, R., and Costantini, R. (2014). Bond returns and market expectations. *Journal of Financial Econometrics*, 12(4):708–729.

Altavilla, C., Giacomini, R., and Ragusa, G. (2017). Anchoring the yield curve using survey expectations. *Journal of Applied Econometrics*, 32(6):1055–1068.

Andersen, T. G. and Benzoni, L. (2010). Do bonds span volatility risk in the u.s. treasury market? a specification test for affine term structure models. *The Journal of Finance*, 65(2):603–653.

Ang, A. and Piazzesi, M. (2003). A no-arbitrage vector autoregression of term structure dynamics with macroeconomic and latent variables. *Journal of Monetary Economics*, 50(4):745–787.

Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4.

Backwell, A. (2021). Unspanned stochastic volatility from an empirical and practical perspective. *Journal of Banking & Finance*, 122:105993.

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

Bauer, M. and Chernov, M. (2024). Interest rate skewness and biased beliefs. *The Journal of Finance*, 79(1):173–217.

Bauer, M. and Rudebusch, G. (2017). Resolving the spanning puzzle in macro-finance term structure models. *Review of Finance*, 21(2):511–553.

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

Bauer, M. D., Pflueger, C. E., and Sunderam, A. (2024). Perceptions about monetary policy. *The Quarterly Journal of Economics*.

Bauer, M. D. and Rudebusch, G. D. (2020). Interest Rates under Falling Stars. *American Economic Review*, 110(5):1316–1354.

Bauer, M. D. and Swanson, E. T. (2023). An alternative explanation for the "fed information effect". *American Economic Review*, 113(3):664–700.

Bernanke, B. S. and Boivin, J. (2003). Monetary policy in a data-rich environment. *Journal of Monetary Economics*, 50(3):525–546.

Bianchi, D., Büchner, M., and Tamoni, A. (2021). Bond risk premiums with machine learning. *Review of Financial Studies*, 34(2):1046–1089.

Borup, D., Eriksen, J. N., Kjær, M. M., and Thyrsgaard, M. (2023). Predicting bond return predictability. *Management Science*.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Caldeira, J., Cordeiro, W., Ruiz, E., and A. P. Santos, A. (2023). Forecasting the yield curve: The role of additional and time-varying decay parameters, conditional heteroscedasticity, and macro-economic factors. *SSRN Electronic Journal*.

Campbell, J. Y. and Shiller, R. J. (1991). Yield Spreads and Interest Rate Movements: A Bird's Eye View. *Review of Economic Studies*, 58(3):495–514.

Carriero, A. and Giacomini, R. (2011). How useful are no-arbitrage restrictions for forecasting the term structure of interest rates? *Journal of Econometrics*, 164(1):21–34.

Carvalho, C., Nechio, F., and Tristão, T. (2021). Taylor rule estimation by ols. *Journal of Monetary Economics*, 124:140–154.

Cieslak, A. (2018). Short-rate expectations and unexpected returns in treasury bonds. *Review of Financial Studies*, 31(9):3265–3306.

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

Cieslak, A. and Vissing-Jorgensen, A. (2021). The economics of the fed put. *The Review of Financial Studies*, 34(9):4045–4089.

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

Collin-Dufresne, P. and Goldstein, R. S. (2002). Do bonds span the fixed income markets? theory and evidence for unspanned stochastic volatility. *The Journal of Finance*, 57(4):1685–1730.

Cooper, I. and Priestley, R. (2009). Time-varying risk premiums and the output gap. *Review of Financial Studies*, 22(7):2601–2633.

Coroneo, L., Giannone, D., and Modugno, M. (2016). Unspanned macroeconomic factors in the yield curve. *Journal of Business & Economic Statistics*, 34(3):472–485.

Dickey, D. A. and Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74(366):427.

Diebold, F. and Mariano, R. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–63.

Diebold, F. X. and Li, C. (2006). Forecasting the term structure of government bond yields. *Journal of Econometrics*, 130(2):337–364.

Diebold, F. X. and Rudebusch, G. D. (2013). *Yield curve modeling and forecasting: the dynamic Nelson-Siegel approach*. The Econometric and Tinbergen Institutes lectures. Princeton University Press, Princeton.

Diebold, F. X., Rudebusch, G. D., and Aruoba, S. B. (2006). The macroeconomy and the yield curve: a dynamic latent factor approach. *Journal of Econometrics*, 131(1-2):309–338.

Duffee, G. (2002). Term premia and interest rate forecasts in affine models. *Journal of Finance*, 57(1):405–443.

Duffee, G. (2013a). Forecasting interest rates. In *Handbook of Economic Forecasting*, pages 385–426. Elsevier.

Duffee, G. R. (2011). Information in (and not in) the Term Structure. *Review of Financial Studies*, 24(9):2895–2934.

Duffee, G. R. (2013b). Bond pricing and the macroeconomy. In *Handbook of the Economics of Finance*, pages 907–967. Elsevier.

Fama, E. and Bliss, R. R. (1987). The information in long-maturity forward rates. *American Economic Review*, 77(4):680–92.

Favero, C. A., Melone, A., and Tamoni, A. (2023). Monetary policy and bond prices with drifting equilibrium rates. *Journal of Financial and Quantitative Analysis*, 59(2):626–651.

Feng, G., Fulop, A., and Li, J. (2022). Real-time macro information and bond return predictability: Does deep learning help? *SSRN Electronic Journal*.

Feng, G., Giglio, S., and Xiu, D. (2020). Taming the factor zoo: A test of new factors. *Journal of Finance*, 75(3):1327–1370.

Fernandes, M. and Vieira, F. (2019). A dynamic nelson–siegel model with forward-looking macroeconomic factors for the yield curve in the US. *Journal of Economic Dynamics and Control*, 106:103720.

Ferreira, M. A. and Santa-Clara, P. (2011). Forecasting stock market returns: The sum of the parts is more than the whole. *Journal of Financial Economics*, 100(3):514–537.

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

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

Giannone, D., Lenza, M., and Primiceri, G. E. (2021). Economic predictions with big data: The illusion of sparsity. *Econometrica*, 89(5):2409–2437.

Giglio, S., Xiu, D., and Zhang, D. (2021). Test assets and weak factors. *NBER Working Paper w29002*.

Goulet-Coulombe, P. (2023). The macroeconomy as a random forest. *Journal of Applied Econometrics*, forthcoming.

Greenwood, R. and Vayanos, D. (2014). Bond supply and excess bond returns. *Review of Financial Studies*, 27(3):663–713.

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

Guidolin, M. and Pedio, M. (2019). Forecasting and trading monetary policy effects on the riskless yield curve with regime switching nelson–siegel models. *Journal of Economic Dynamics and Control*, 107(C).

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

Gürkaynak, R. S. and Wright, J. H. (2012). Macroeconomics and the term structure. *Journal of Economic Literature*, 50(2):331–367.

Gürkaynak, R. S., Sack, B., and Swanson, E. (2005). Do Actions Speak Louder Than Words? The Response of Asset Prices to Monetary Policy Actions and Statements. *International Journal of Central Banking*, 1(1).

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer New York.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Huang, J.-Z. and Shi, Z. (2023). Machine-learning-based return predictors and the spanning controversy in macro-finance. *Management Science*, 69(3):1780–1804.

Hännikäinen, J. (2017). When does the yield curve contain predictive power? evidence from a data-rich environment. *International Journal of Forecasting*, 33(4):1044–1064.

Jia, P., Shen, H., and Zheng, S. (2023). Monetary policy rules and opinionated markets. *Economics Letters*, 223:110995.

Joslin, S., Priebsch, M., and Singleton, K. (2014). Risk premiums in dynamic term structure models with unspanned macro risks. *Journal of Finance*, 69(3):1197–1233.

Kuttner, K. N. (2018). Outside the box: Unconventional monetary policy in the great recession and beyond. *Journal of Economic Perspectives*, 32(4):121–146.

Ledoit, O. and Wolf, M. (2008). Robust performance hypothesis testing with the sharpe ratio. *Journal of Empirical Finance*, 15(5):850–859.

Li, H. and Zhao, F. (2006). Unspanned stochastic volatility: Evidence from hedging interest rate derivatives. *The Journal of Finance*, 61(1):341–378.

Li, H. and Zhao, F. (2009). Nonparametric estimation of state-price densities implicit in interest rate cap prices. *Review of Financial Studies*, 22(11):4335–4376.

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

Liu, Y. and Wu, C. (2021). Reconstructing the yield curve. *Journal of Financial Economics*, 142(3):1395–1425.

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

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

Medeiros, M. C., Vasconcelos, G. F. R., Álvaro Veiga, and Zilberman, E. (2021). Forecasting inflation in a data-rich environment: The benefits of machine learning methods. *Journal of Business & Economic Statistics*, 39(1):98–119.

Moench, E. (2008). Forecasting the yield curve in a data-rich environment: A no-arbitrage factor-augmented var approach. *Journal of Econometrics*, 146(1):26–43.

Nelson, C. and Siegel, A. F. (1987). Parsimonious modeling of yield curves. *Journal of Business*, 60(4):473–89.

Newey, W. K. and West, K. D. (1987). A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica*, 55(3):703–708.

Piazzesi, M., Salomao, J., and Schneider, M. (2015). Trend and cycle in bond premia. *Unpublished Manuscript*.

Riva, R. (2024). How much unspanned volatility can different shocks explain? *SSRN*.

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

Schmeling, M., Schrimpf, A., and Steffensen, S. A. (2022). Monetary policy expectation errors. *Journal of Financial Economics*, 146(3):841–858.

Sihvonen, M. (2024). Yield curve momentum. *Review of Finance*, 28(3):805–830.

Stock, J. H. and Watson, M. W. (2002a). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97(460):1167–1179.

Stock, J. H. and Watson, M. W. (2002b). Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics*, 20(2):147–162.

Swanson, E. T. (2023). The importance of fed chair speeches as a monetary policy tool. *AEA Papers and Proceedings*, 113:394–400.

Swanson, E. T. and Jayawickrema, V. (2024). Speeches by the fed chair are more important than fomc announcements: An improved high-frequency measure of u.s. monetary policy shocks. *Unpublished Manuscript*.

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

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

van Dijk, D., Koopman, S. J., van der Wel, M., and Wright, J. H. (2013). Forecasting interest rates with shifting endpoints. *Journal of Applied Econometrics*, 29(5):693–712.

Welch, I. and Goyal, A. (2007). A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies*, 21(4):1455–1508.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67(2):301–320.

Supplementary Appendix

A Additional Figures

Figure A.1: Spectral decomposition of the FRED-MD data set. We normalize each of the variables and compute the eigenvalues of the correlation matrix. We show how much of the total variation is commanded by each eigenvector, denoted by the relative size of the corresponding eigenvalue.

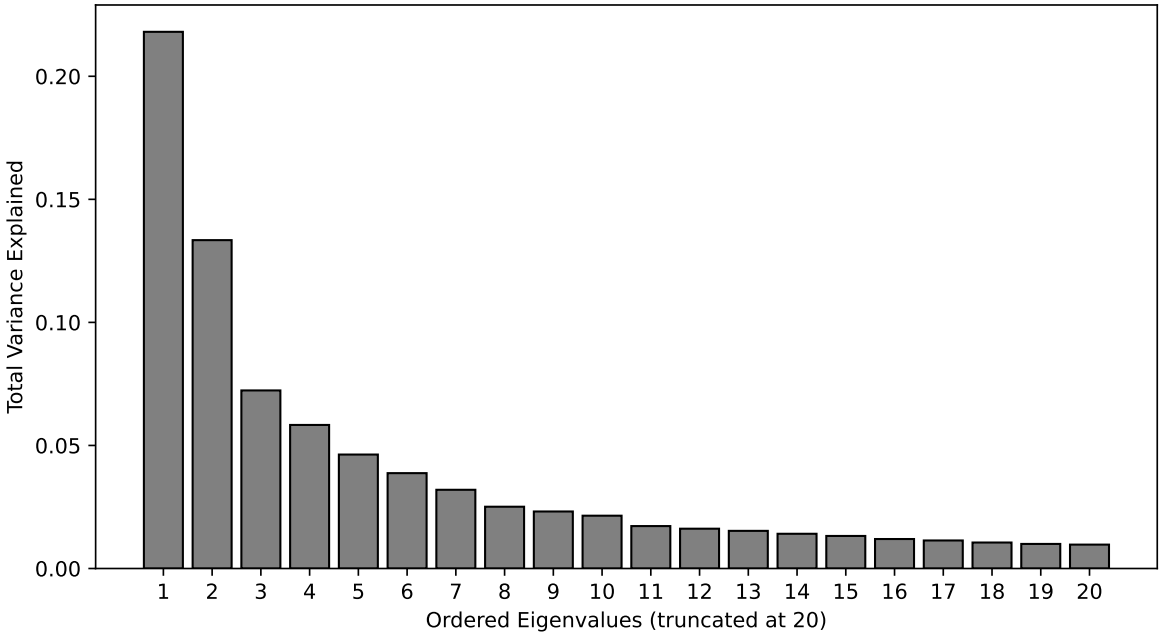


Figure A.2: Relative MSE predicting returns using three principal components of the yield curve as controls. For each maturity, we show the ratio between the MSE attained with different numbers of principal components from the macroeconomic data and the baseline model that uses information only from the yield curve itself. The sample for maturity of less than 120 months ranges from 1973 to 2021, while it starts in 1985 for the other maturities. For any of the maturities, the out-of-sample period starts in January 1990. We use the linear model in (2) to make the forecasts.

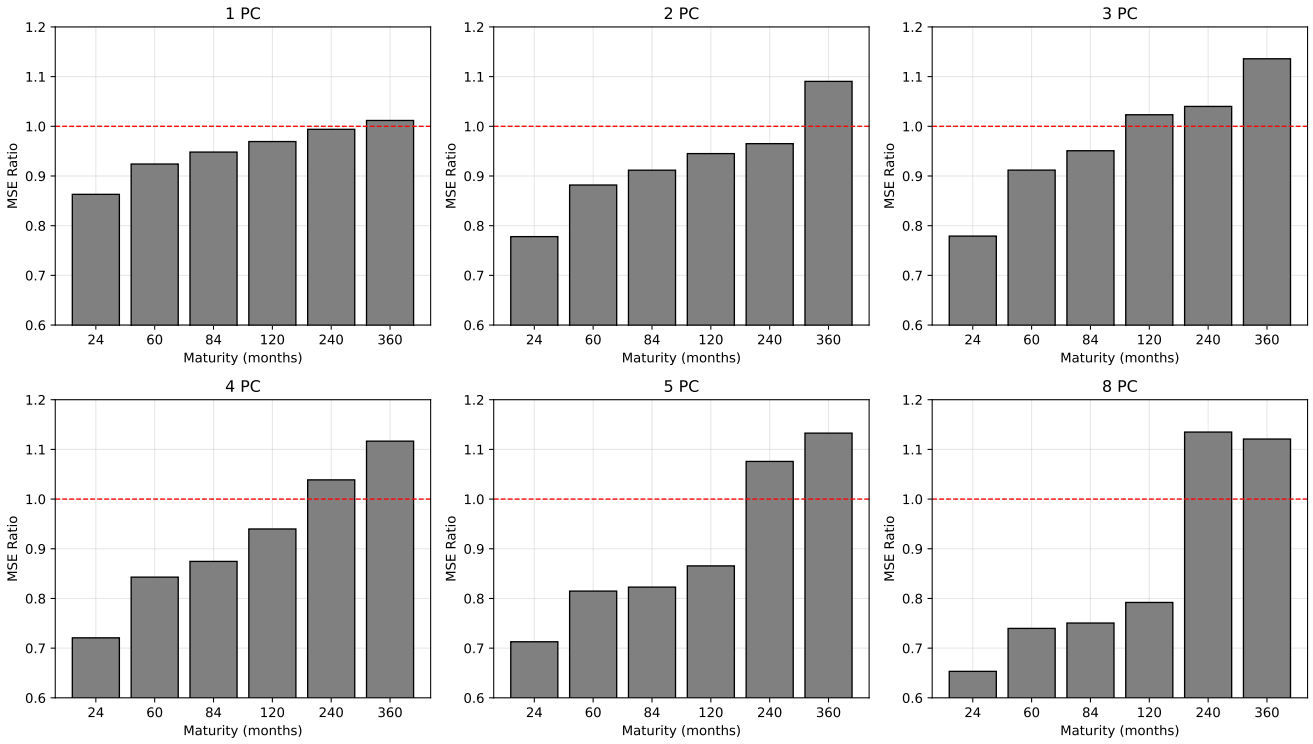


Figure A.3: Relative MSE predicting bond returns using the forward rates as controls and using vintage data from the FRED-MD website. To deal with the ragged edge problem, we use the EM algorithm method outlined in Stock and Watson (2002b) and recently used in Sihvonen (2024). See the discussion in Figure 1.

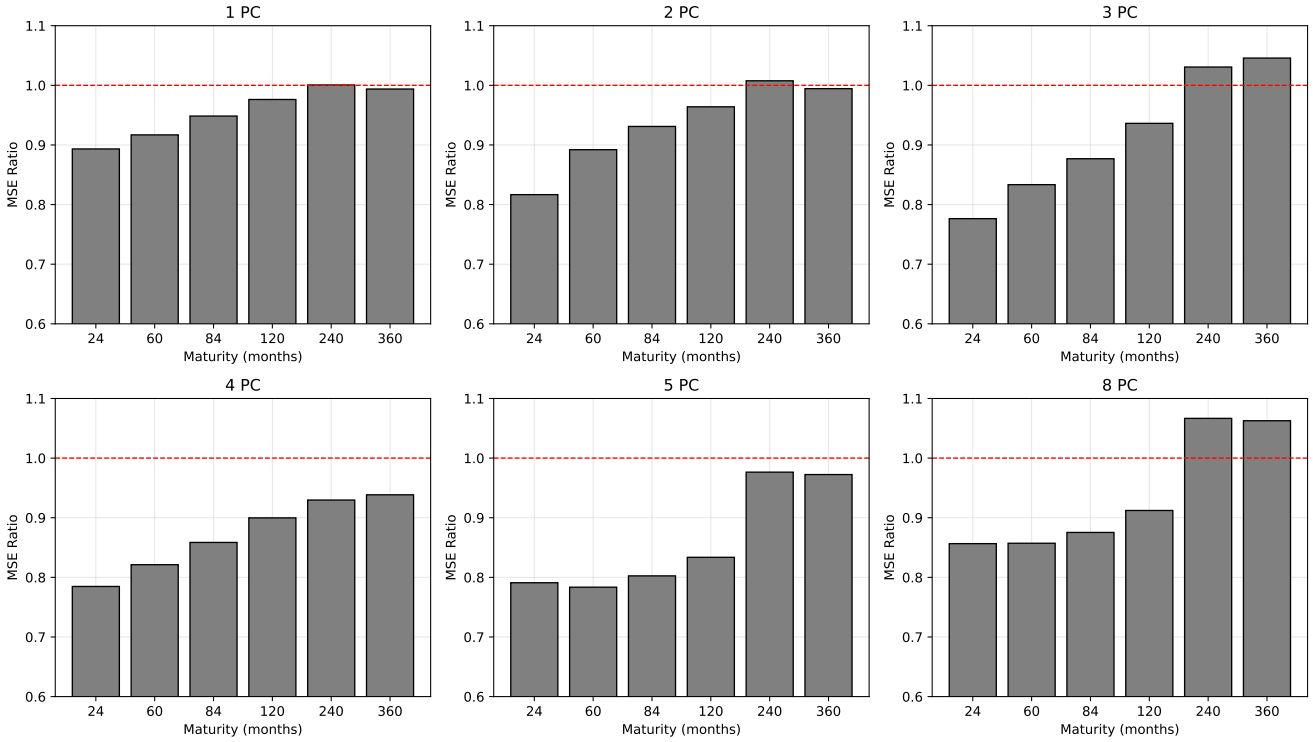


Figure A.4: This figure reports the Sharpe ratio improvement the mean-variance consumer can achieve when using principal components of the FRED-MD data set to trade bonds of different maturities. The out-of-sample period is 1990-2021. The dots represent point estimates while the gray bars represent 95% confidence intervals using the asymptotic framework from Ledoit and Wolf (2008). This plot focuses on the strict case, i.e., assuming no leverage and a short-selling constraint.

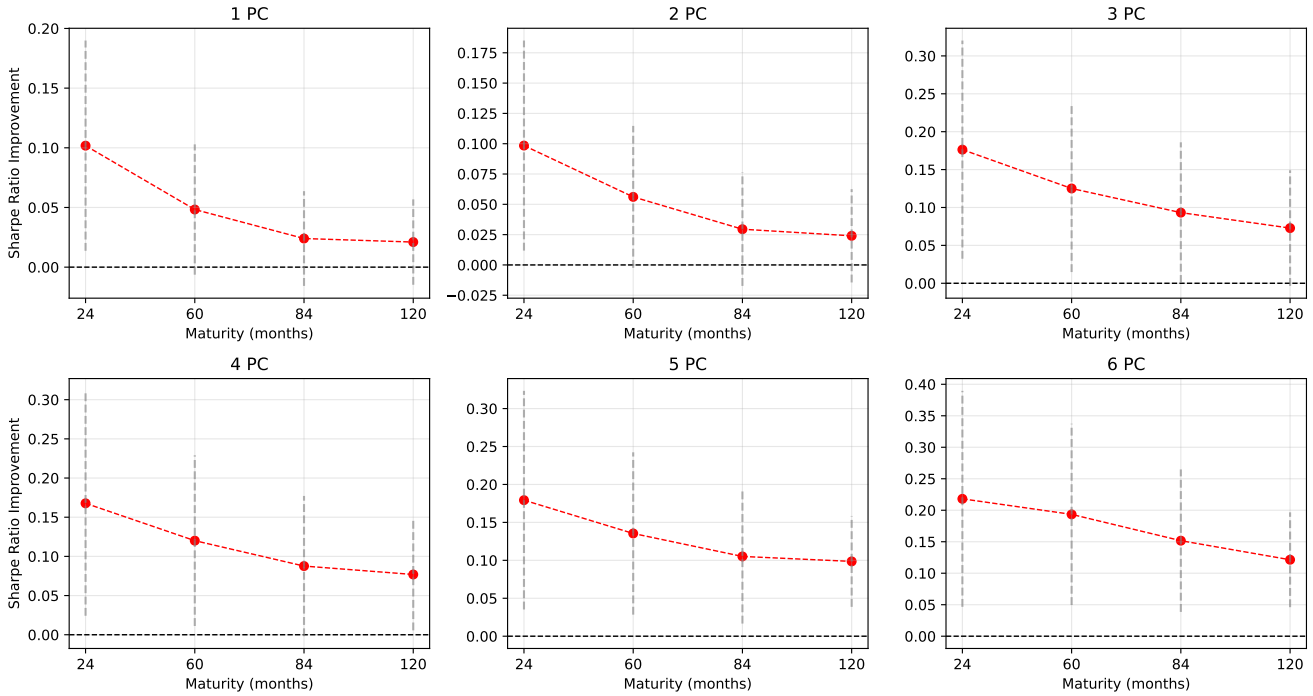


Figure A.5: This figure reports the Sharpe ratio improvement the mean-variance consumer can achieve when using principal components of the FRED-MD dataset to trade bonds of different maturities. The out-of-sample period is 1990-2021. The dots represent point estimates while the gray bars represent 95% confidence intervals using the asymptotic framework from Ledoit and Wolf (2008). This plot focuses on the more general case when we assume that $-1 \leq \boldsymbol{w}_t \leq 2$.

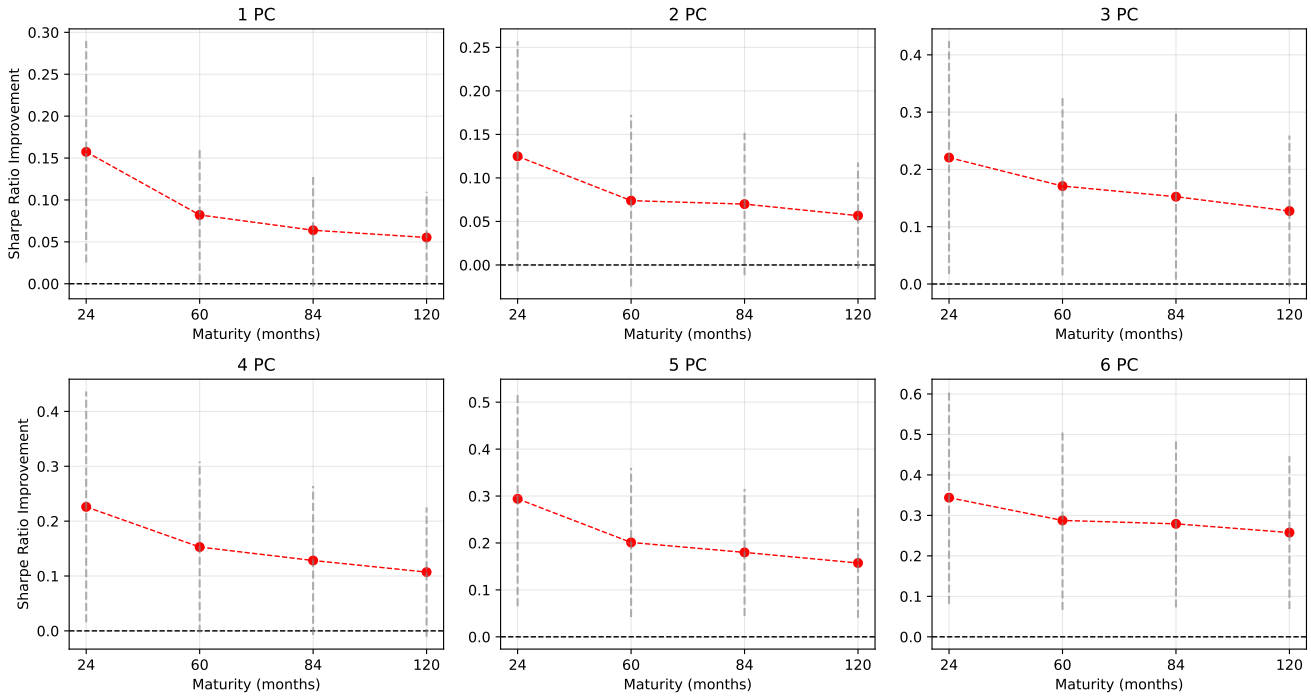


Figure A.6: Forecasts for β_1 both targeting its level (left) and its innovation (right) created by the ElasticNet. The out-of-sample period is January 1990 to December 2021. The black dashed line represents the realizations we estimated as shown in Figure 3. The forecasts for the level show a large gap with respect to the black dashed line, indicating that these forecasts are consistently overestimating the target, leading to poor out-of-sample performance - both with and without macroeconomic data. This happens because the realizations of β_1 in the in-sample period were generally higher than in the out-of-sample one (see Figure 3). If the estimator could set a coefficient close to 1 for the lagged value of β_1, for example, the forecasts would slowly adjust to lower realizations of the long-run factor. But the penalization terms make this a costly choice. When we predict innovations, we have a very different picture. Since the target we have is stationary, we have no problem forecasting it with information from the yield curve plus stationary variables from the FRED-MD data set. Crucially, both solid lines in the panels on the left are on top of each other, indicating that the addition of macroeconomic variables did not enhance the forecast for the given target.

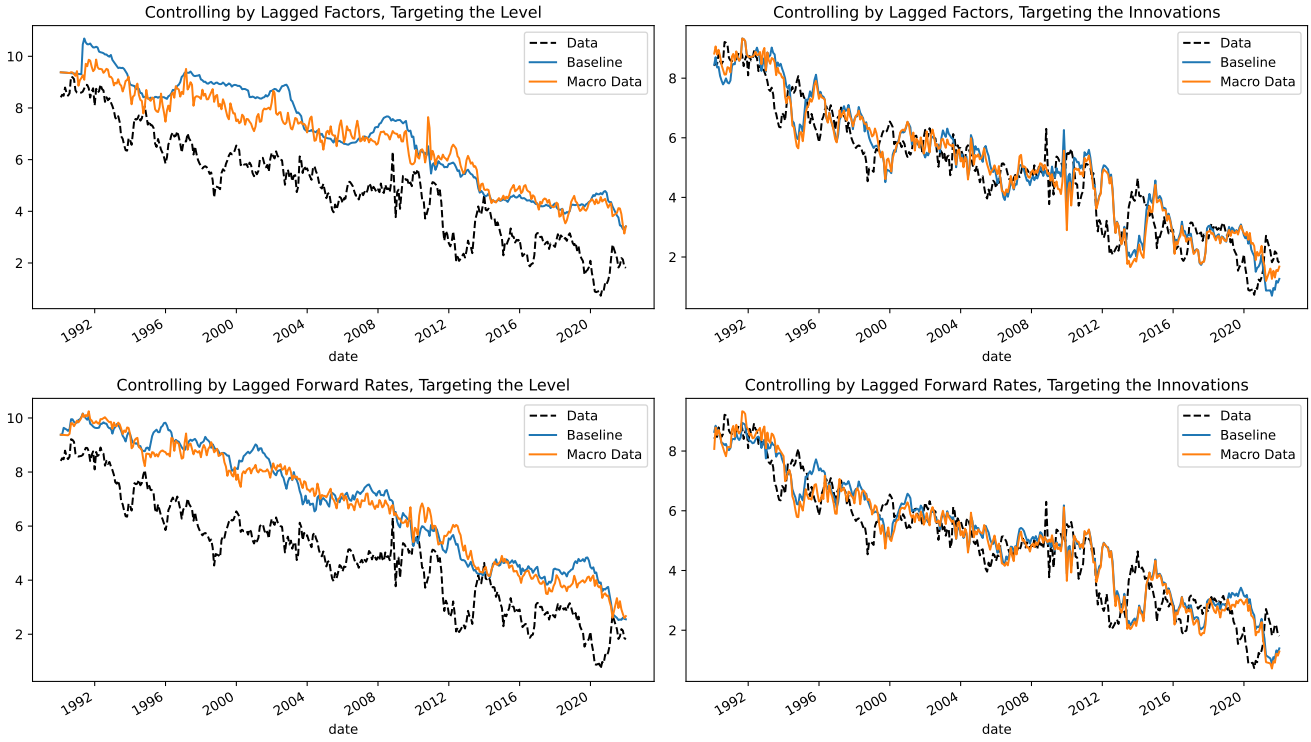


Figure A.7: Individual feature importance for each variable used in splits for the Random Forest strategy. We compute the feature importance of each variable for each tree and then average over trees and over time, normalizing everything at the end. Blue bars represent variables from the FRED-MD dataset, while the red ones are information extracted from the yield curve alone. The out-of-sample period is 1990-2021, and we used 500 trees for each forecast.

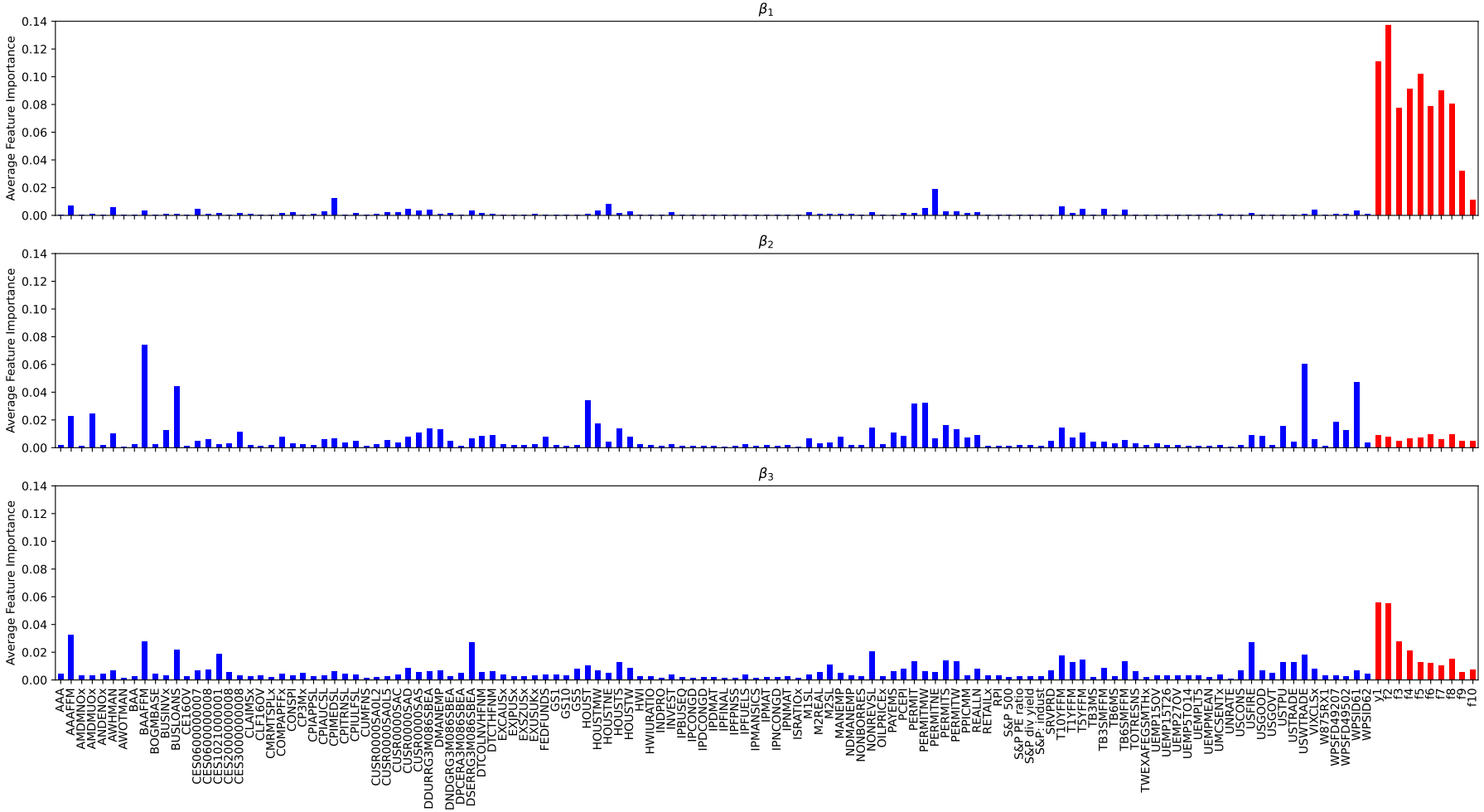


Figure A.8: We plot the decomposition of yields (blue) into the ES_t term (orange) and the TP_t term (black) following Adrian et al. (2013). The black line is the gap between the other two. The left panel shows the decomposition for the 2-year rate, while the one on the right shows the decomposition for the 10-year maturity.

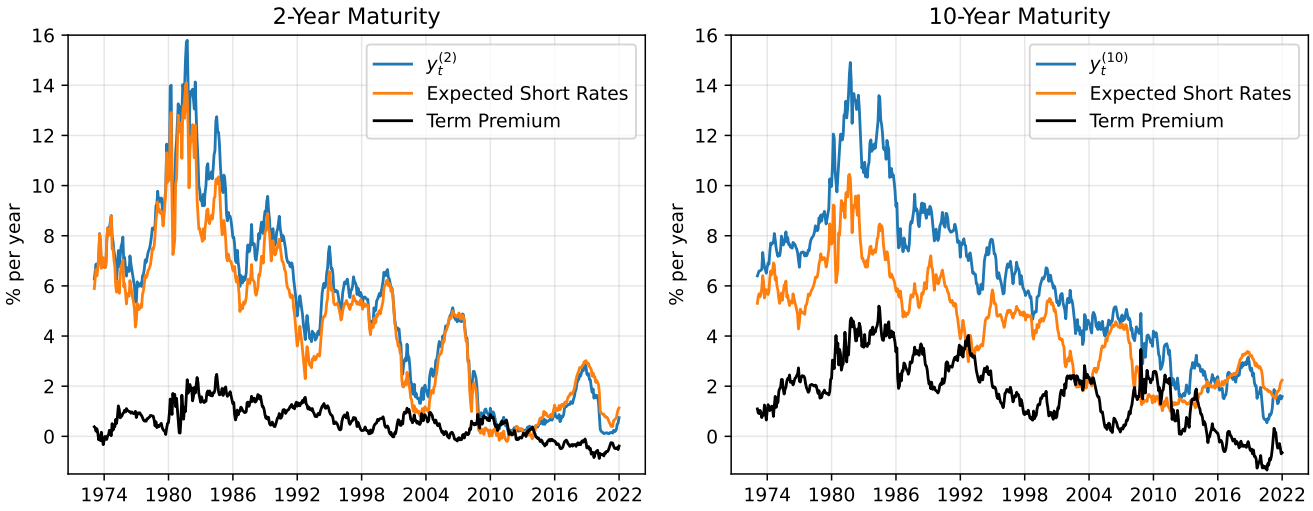


Figure A.9: For each date in the 1973-2021 range and each maturity, we compute the ratio between ES_t and y_t and plot the average ratio over time.

Figure A.10: The left panel shows different versions of the Taylor Rule from (15). Different rules use different inflation measures. The panel on the right shows the gap between the Fed Funds rate and the fitted values.

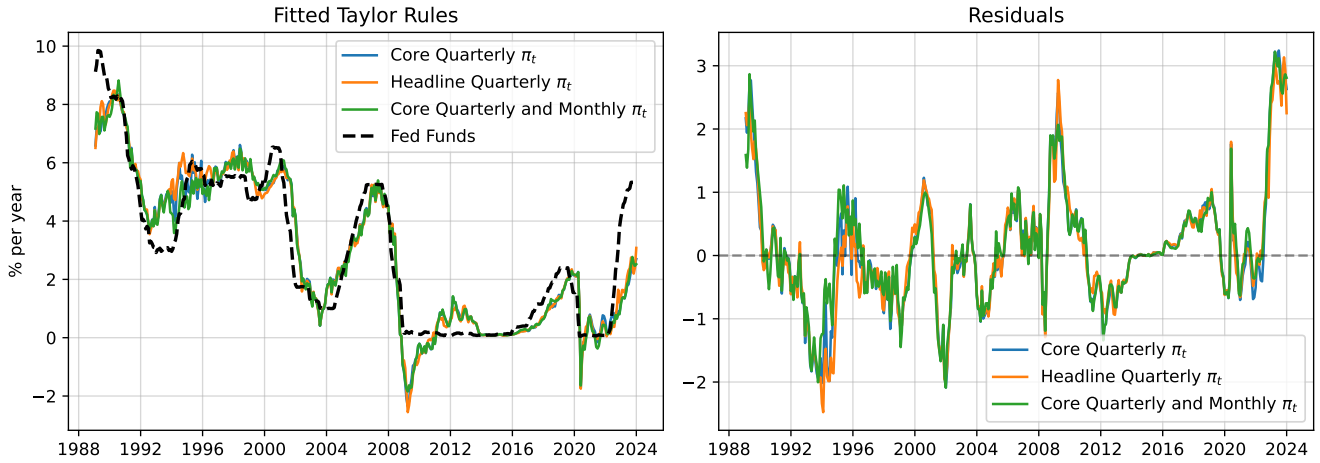


Figure A.11: Point estimates for b from (16) with 95% confidence intervals. Different columns use deviations from different Taylor Rules, which control for different inflation measures. The top row uses loss functions of forecasts that use forward rates to control for the information already in the yield curve. The bottom row uses lagged Nelson-Siegel factors as controls.

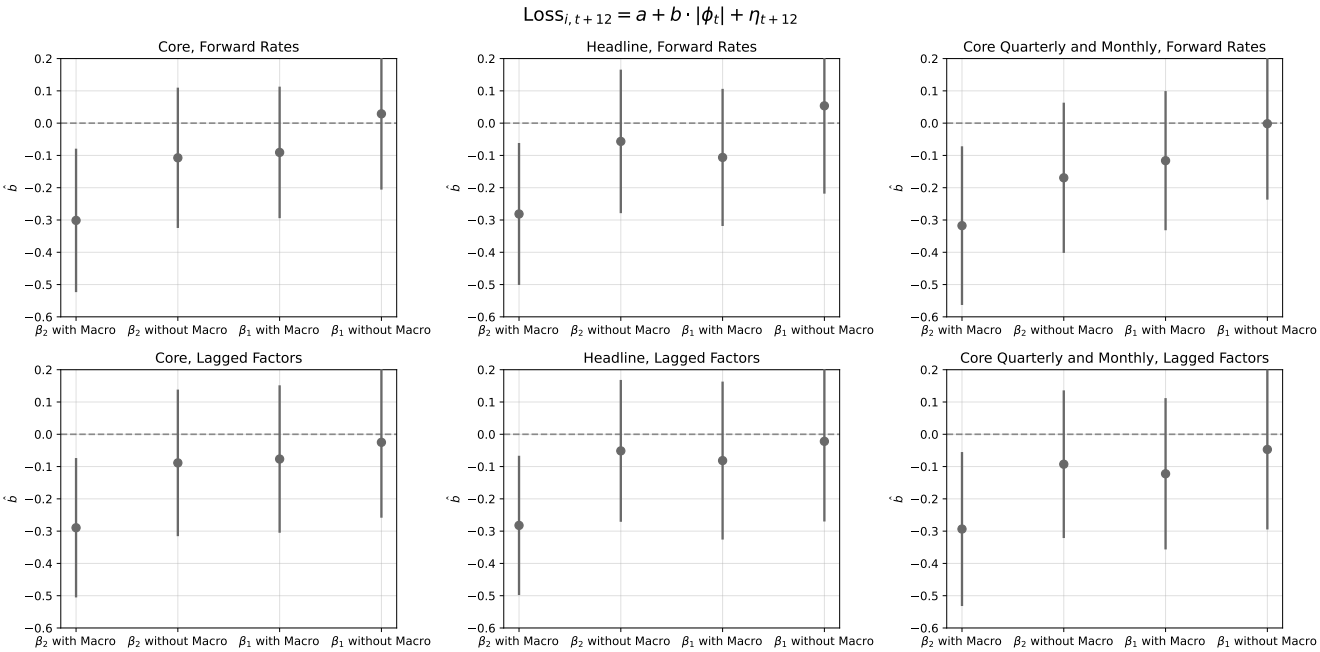
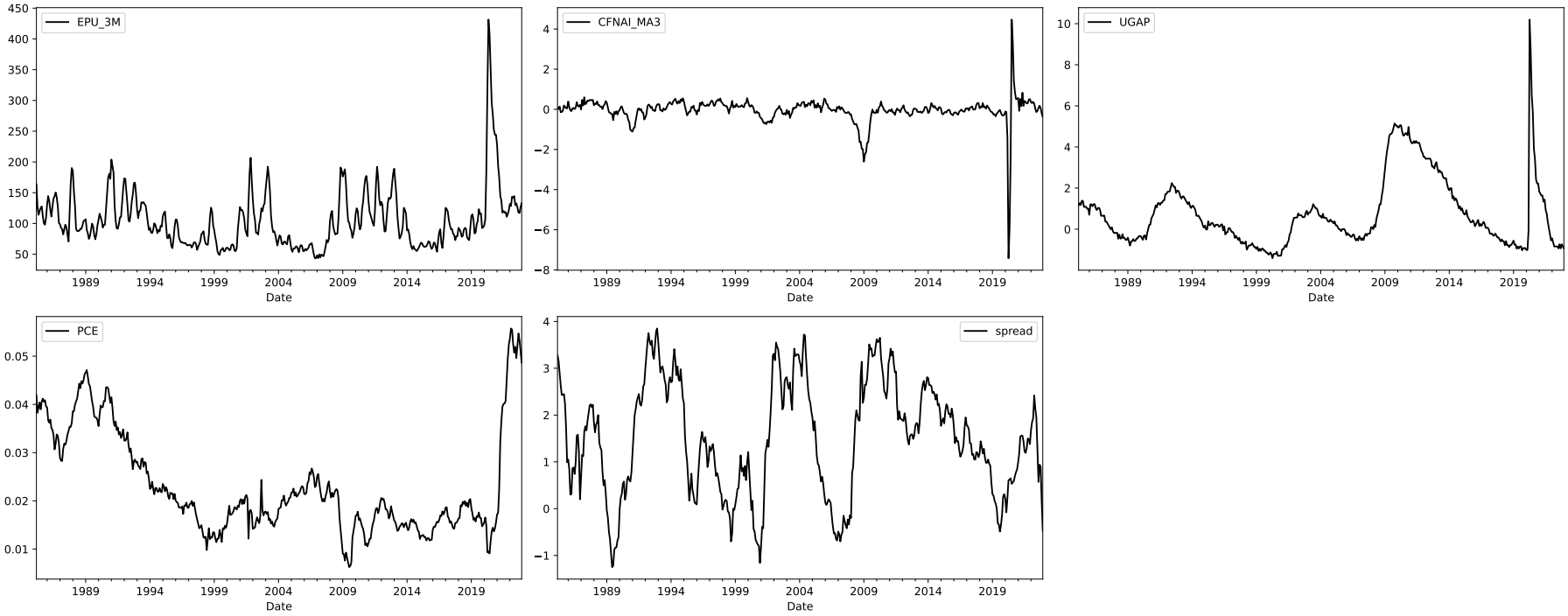


Figure A.12: This figure plots the evolution of the variables used in the conditional predictive ability test. The out-of-sample period is 1990-2021.



B Additional Tables

B.1 In-sample Evidence

Table B.1 presents results for the predictive regression in (2) estimated over the whole sample (i.e., in-sample) when we choose $d = 3$ and let C_t control the first three principal components of the yield curve. Different groups of columns let the maturity n increase from left to right, while different regressions for the same maturity allow for a greater number of principal components extracted from the FRED-MD data set. The sample size for both the 30-year and the 20-year maturities is reduced since these longer maturities started being traded later than 1973 when we started our analysis. In that case, we use the period 1985-2021. We also append at the last row the adjusted R^2 from a regression that imposes $\gamma_n = 0$, so the same value is repeated for each maturity. We omit estimates for θ_n for the sake of space.

The relative gain from the addition of macroeconomic data measured as the increase in the adjusted R^2 from the baseline case to the most complete specification is stronger for the 2-year maturity than for the other ones. In that case, it increases more than threefold (from 12% to 40%). However, its increase is less expressive for longer maturities. For example, for the 20-year maturity, it increases from 14% to 23%. We confirm that the same type of pattern arises when C_t contains the risk-free rate and forward rates, as displayed in Table B.2 below. When controlling for the forward rates, we actually find less evidence of the statistical significance of γ_n for the longer maturities. For example, the coefficient on the first principal component is not significant anymore.

Tables B.3 and B.4 report regression results for the same type of exercise but use the estimated Nelson-Siegel factors as dependent variables. The increase in R^2 with the addition of macroeconomic data is marginal for β_1 and sizable for β_2.

Table B.1: In-sample predictive regression (2) of excess bond returns on principal components of macro data controlling for the first three principal components of the yield curve. Only estimates of γ_n are reported. Standard errors are computed using Newey and West (1987). The sample for the first two columns goes from 1973 to 2021, while it starts in 1985 for the two last ones, which is when the 30-year yield data becomes available in the data set provided by Liu and Wu (2021). Stars denote significance at 10%, 5%, and 1%, respectively.

	2-year			10-year			20-year			30-year		
PC 1	0.09***	0.12***	0.13***	0.04**	0.06***	0.08***	-0.05***	-0.04**	-0.04**	-0.06***	-0.06**	-0.06**
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)	(0.02)
PC 2		-0.06**	-0.08**		-0.05	-0.07*		0.04	0.04		0.02	0.01
		(0.03)	(0.04)		(0.03)	(0.04)		(0.03)	(0.03)		(0.03)	(0.04)
PC 3		0.10***	0.11***		0.06**	0.08***		0.02	0.02		0.01	-0.00
		(0.02)	(0.02)		(0.03)	(0.02)		(0.03)	(0.03)		(0.04)	(0.03)
PC 4			-0.04			-0.08***			-0.05**			-0.07**
			(0.03)			(0.02)			(0.02)			(0.03)
PC 5			-0.07**			-0.11***			-0.04			-0.10***
			(0.03)			(0.03)			(0.03)			(0.04)
PC 6			0.05			0.09***			0.07**			0.06
			(0.03)			(0.03)			(0.03)			(0.05)
PC 7			0.06*			0.02			-0.03			-0.03
			(0.03)			(0.02)			(0.03)			(0.03)
PC 8			-0.09**			-0.09***			-0.03			-0.06
			(0.03)			(0.03)			(0.03)			(0.04)
N	588	588	588	588	588	588	422	422	422	422	422	422
R^2 Adj.	0.25	0.33	0.40	0.20	0.23	0.36	0.18	0.20	0.23	0.16	0.16	0.24
R^2 Adj. (No Macro Data)	0.12	0.12	0.12	0.17	0.17	0.17	0.14	0.14	0.14	0.12	0.12	0.12

Table B.2: In-sample predictive regression of excess bond returns on principal components of macro data controlling by the forward rates, as in equation (2). We only report estimates of γ_n. See the discussion about Table B.1 in the main text. Standard errors are computed using Newey and West (1987). The sample for the first two columns goes from 1973 to 2021, while it starts in 1985 for the two last ones, which is when the 30-year yield data becomes available in the data set provided by Liu and Wu (2021).

	2-year			10-year			20-year			30-year		
PC 1	0.09***	0.12***	0.13***	0.04**	0.07***	0.07***	-0.01	-0.00	0.00	-0.03	-0.02	-0.03
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)	(0.02)	(0.03)	(0.04)
PC 2		-0.07**	-0.07**		-0.07***	-0.06**		-0.01	0.00		0.00	0.02
		(0.03)	(0.03)		(0.02)	(0.02)		(0.04)	(0.05)		(0.05)	(0.06)
PC 3		0.11***	0.11***		0.08***	0.08***		0.05**	0.05*		0.04	0.03
		(0.03)	(0.02)		(0.03)	(0.02)		(0.03)	(0.03)		(0.03)	(0.03)
PC 4		-0.02	-0.02		-0.05***	-0.06***		-0.06***	-0.06***		-0.09***	-0.08***
		(0.02)	(0.03)		(0.02)	(0.02)		(0.02)	(0.02)		(0.02)	(0.02)
PC 5		-0.04	-0.04		-0.09***	-0.08***		-0.08**	-0.08*		-0.09**	-0.09*
		(0.03)	(0.03)		(0.03)	(0.03)		(0.04)	(0.05)		(0.05)	(0.05)
PC 6			0.03			0.07***			0.04			0.06
			(0.03)			(0.03)			(0.04)			(0.05)
PC 7			0.06*			0.04			0.01			0.01
			(0.03)			(0.03)			(0.03)			(0.03)
PC 8			-0.08***			-0.08***			-0.04			-0.04
			(0.03)			(0.03)			(0.04)			(0.05)
N	588	588	588	588	588	588	422	422	422	422	422	422
R^2 Adj.	0.28	0.36	0.40	0.28	0.36	0.40	0.16	0.23	0.24	0.15	0.22	0.23
R^2 Adj. (No Macro Data)	0.15	0.15	0.15	0.25	0.25	0.25	0.16	0.16	0.16	0.14	0.14	0.14

Table B.3: In-sample predictive regressions targeting the level of the factors as in (8). The baseline model uses the lagged Nelson-Siegel factors as controls for the yield curve. We only show estimates for γ_i. C_t stores the lagged values of the Nelson-Siegel factors. Standard errors are compute using Newey and West (1987). The two last rows report the adjust R^2 when we set $\gamma_i = 0$ ("No Macro"). We use data from 1973 until 2021. Stars denote significance at 10%, 5%, and 1%, respectively.

	β_1 (1)	β_1 (2)	β_1 (3)	β_1 (4)	β_2 (1)	β_2 (2)	β_2 (3)	β_2 (4)	β_3 (1)	β_3 (2)	β_3 (3)	β_3 (4)
PC 1	-0.04	-0.06**	-0.09***	-0.06**	-0.21***	-0.19***	-0.16***	-0.17***	-0.17***	-0.17***	-0.21***	-0.25***
	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)	(0.05)	(0.05)	(0.05)	(0.04)	(0.05)	(0.05)	(0.05)
PC 2		0.02	0.04	0.01		-0.11**	-0.11**	-0.12**		-0.09*	-0.07	-0.05
		(0.03)	(0.03)	(0.02)		(0.05)	(0.05)	(0.06)		(0.05)	(0.04)	(0.04)
PC 3		0.05*	0.07**	0.06**		0.03	0.02	0.03		0.13***	0.15***	0.17***
		(0.03)	(0.03)	(0.03)		(0.05)	(0.05)	(0.05)		(0.05)	(0.05)	(0.05)
PC 4			-0.06***	-0.09***			-0.01	-0.01			0.04	0.07*
			(0.02)	(0.02)			(0.04)	(0.04)			(0.04)	(0.04)
PC 5			0.15***	0.14***			-0.08	-0.08			0.12**	0.13***
			(0.03)	(0.02)			(0.05)	(0.05)			(0.05)	(0.05)
PC 6				-0.17***				0.03				0.13*
				(0.03)				(0.07)				(0.07)
PC 7				0.01				-0.03				-0.14***
				(0.03)				(0.06)				(0.05)
PC 8				-0.02				-0.21***				-0.12
				(0.04)				(0.07)				(0.09)
N	588	588	588	588	588	588	588	588	588	588	588	588
R^2 Adj.	0.89	0.89	0.90	0.91	0.42	0.45	0.45	0.48	0.50	0.53	0.53	0.55
R^2 Adj. (No Macro)	0.89	0.89	0.89	0.89	0.33	0.33	0.33	0.33	0.47	0.47	0.47	0.47

Table B.4: In-sample predictive regressions targeting the level of the factors as in (8). The baseline model uses the forward rates as controls for the yield curve. We only show estimates for γ_i. C_t stores the lagged values of the Nelson-Siegel factors. Standard errors are computed using Newey and West (1987). The two last rows report the adjusted R^2 when we set $\gamma_i = 0$ ("No Macro"). We use data from 1973 until 2021. Stars denote significance at 10%, 5%, and 1%, respectively.

	β_1 (1)	β_1 (2)	β_1 (3)	β_1 (4)	β_2 (1)	β_2 (2)	β_2 (3)	β_2 (4)	β_3 (1)	β_3 (2)	β_3 (3)	β_3 (4)
PC 1	-0.06**	-0.09***	-0.13***	-0.10***	-0.23***	-0.21***	-0.19***	-0.21***	-0.15***	-0.15***	-0.17***	-0.19***
	(0.02)	(0.03)	(0.03)	(0.03)	(0.04)	(0.04)	(0.05)	(0.05)	(0.04)	(0.05)	(0.05)	(0.06)
PC 2		0.05**	0.07***	0.05**		-0.12***	-0.13***	-0.12**		-0.10***	-0.09**	-0.09**
		(0.02)	(0.02)	(0.02)		(0.05)	(0.05)	(0.05)		(0.04)	(0.04)	(0.04)
PC 3		0.07**	0.10***	0.09***		0.07	0.06	0.06		0.10**	0.12**	0.12**
		(0.03)	(0.03)	(0.03)		(0.05)	(0.05)	(0.05)		(0.05)	(0.05)	(0.05)
PC 4			-0.06***	-0.09***			0.02	0.03			0.05	0.06
			(0.02)	(0.02)			(0.05)	(0.05)			(0.04)	(0.04)
PC 5			0.18***	0.17***			-0.07	-0.06			0.05	0.05
			(0.02)	(0.02)			(0.05)	(0.05)			(0.06)	(0.05)
PC 6				-0.13***				0.07				0.05
				(0.03)				(0.07)				(0.06)
PC 7				-0.01				-0.06				-0.06
				(0.03)				(0.05)				(0.05)
PC 8				-0.02				-0.16***				-0.18**
				(0.03)				(0.06)				(0.09)
N	588	588	588	588	588	588	588	588	588	588	588	588
R^2 Adj.	0.90	0.91	0.92	0.93	0.47	0.51	0.51	0.53	0.54	0.56	0.56	0.57
R^2 Adj. (No Macro)	0.90	0.90	0.90	0.90	0.36	0.36	0.36	0.36	0.52	0.52	0.52	0.52

B.2 Additional Out-of-sample Evidence

Table B.5: We report the out-of-sample R^2 attained from the model in (8) for the baseline with no macroeconomic data included and with different numbers of principal components. Negative values imply we couldn't beat a random walk. We also show p-values to compare whether any improvement was statistically significant, comparing the "No Macro" baseline and the different forecasts. Here we control for the lagged Nelson-Siegel factors, while we controlled for the forward rates in the main text.

		Number of Macro PCs						p-values					
Target	No Macro	1	2	3	4	5	8	1	2	3	4	5	8
β_1	-0.10	-0.10	-0.11	-0.14	-0.11	-0.07	0.06	0.51	0.67	0.83	0.56	0.36	0.04
β_2	0.06	0.07	0.21	0.20	0.20	0.20	0.17	0.31	0.01	0.15	0.16	0.18	0.28
β_3	-0.11	-0.14	-0.06	-0.05	-0.05	-0.06	-0.08	0.89	0.16	0.19	0.20	0.23	0.39

Table B.6: We report the out-of-sample R^2 from the forecasting model in (17). The out-of-sample period is 1990-2021. We set $r = 2$ and target both levels and innovations, controlling either the forward rates or the lagged Nelson-Siegel factors. The p-values assess whether improvements over "No Macro" were statistically significant.

Panel A: Controlling for Forward Rates, $r = 2$							
	Number of Macro PCs				p-values		
Target	No Macro	3	5	8	3	5	8
β_1	-0.21	-0.16	-0.17	-0.02	0.25	0.36	0.03
β_2	-0.08	0.20	0.19	0.28	0.02	0.03	0.02
β_3	-0.12	-0.08	-0.10	-0.15	0.27	0.37	0.62
$\Delta\beta_1$	-0.19	-0.13	-0.14	-0.01	0.20	0.32	0.03
$\Delta\beta_2$	-0.11	0.18	0.18	0.26	0.02	0.02	0.02
$\Delta\beta_3$	-0.10	-0.06	-0.09	-0.11	0.29	0.44	0.55
Panel B: Controlling for Lagged Betas, $r = 2$							
	Number of Macro PCs				p-values		
Target	No Macro	3	5	8	3	5	8
β_1	-0.10	-0.12	-0.13	0.00	0.67	0.65	0.11
β_2	0.06	0.20	0.20	0.25	0.07	0.07	0.08
β_3	-0.11	-0.06	-0.09	-0.12	0.16	0.37	0.57
$\Delta\beta_1$	-0.10	-0.12	-0.13	0.00	0.67	0.65	0.11
$\Delta\beta_2$	0.06	0.20	0.20	0.25	0.07	0.07	0.08
$\Delta\beta_3$	-0.11	-0.06	-0.09	-0.12	0.16	0.37	0.57

Table B.7: We report the out-of-sample R^2 from the forecasting model in (17). The out-of-sample period is 1990-2021. We set $r = 3$ and target both levels and innovations, controlling either for the forward rates or the lagged Nelson-Siegel factors. The p-values assess whether improvements over "No Macro" were statistically significant.

Panel A: Controlling for Forward Rates, $r = 3$							
	Number of Macro PCs				p-values		
Target	No Macro	3	5	8	3	5	8
β_1	-0.21	-0.15	-0.13	-0.04	0.19	0.20	0.04
β_2	-0.08	0.13	0.09	0.03	0.09	0.15	0.30
β_3	-0.12	-0.10	-0.13	-0.19	0.35	0.56	0.81
$\Delta\beta_1$	-0.19	-0.12	-0.11	-0.02	0.15	0.17	0.03
$\Delta\beta_2$	-0.11	0.12	0.08	0.03	0.08	0.13	0.25
$\Delta\beta_3$	-0.10	-0.08	-0.12	-0.20	0.41	0.66	0.86
Panel B: Controlling for Lagged Betas $r = 3$							
	Number of Macro PCs				p-values		
Target	No Macro	3	5	8	3	5	8
β_1	-0.10	-0.12	-0.09	-0.02	0.67	0.47	0.14
β_2	0.06	0.13	0.11	0.05	0.33	0.37	0.53
β_3	-0.11	-0.09	-0.14	-0.20	0.40	0.70	0.83
$\Delta\beta_1$	-0.10	-0.12	-0.09	-0.02	0.67	0.47	0.14
$\Delta\beta_2$	0.06	0.13	0.11	0.05	0.33	0.37	0.53
$\Delta\beta_3$	-0.11	-0.09	-0.14	-0.20	0.40	0.70	0.83

Table B.8: We report out-of-sample R^2 using the Random Forest method with 500 trees estimated at each point in time. The out-of-sample period is 1990-2021. Negative values mean we couldn't beat a random walk. We use 120 months for the rolling window. The p-values assess whether any improvement over the "No Macro" baseline is significant.

	Lagged Betas			Forward Rates		
Target	No Macro	All Macro	p-value	No Macro	All Macro	p-value
β_1	-0.87	-1.01	0.74	-0.63	-0.66	0.58
β_2	-0.07	0.15	0.03	-0.20	0.08	0.01
β_3	-0.55	-0.19	0.01	-0.70	-0.23	0.00

C Alternative Estimation Procedures

Now, we briefly discuss different estimation procedures. Another natural way of estimating (4) is using non-linear least squares (NLS) allowing λ_t to be estimated period by period. This alternative approach, in principle, can do no worse fitting the yield curve than our method since it has one extra parameter to be estimated. In practice, however, we have to set up one numerical optimization scheme for each date, and there is no guarantee of convergence towards a global solution at a given point in time. We experimented with this approach and found that whenever the numerical optimization converged, estimated factors were very close to the ones found by OLS. Nevertheless, the numerical optimization would not converge for roughly 8% of the dates considered. In these cases, the values attained by the factors were extreme. Since we ultimately seek to forecast these factors, these extreme realizations would generate artificially large forecast errors that could invalidate our posterior analyses due to numerical instabilities.

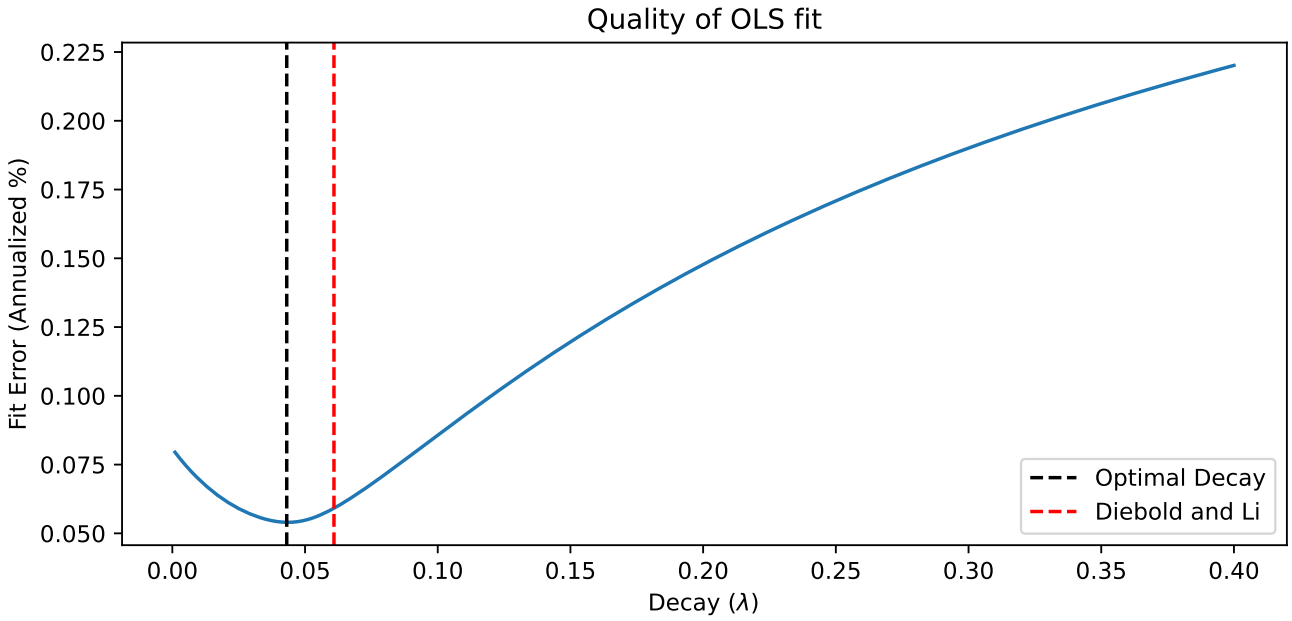
A second possible way to estimate factors and the decay parameter λ is using a two-step approach. For each given value of λ, we can estimate factors for all dates using the estimator in (7) and compute, for example, a time series of the sum of squared residuals in the cross-section of yields. The average of this time series can be understood as a measure of goodness-of-fit for the particular constant value of λ considered. An optimal value of λ in this sense is the one that minimizes this average error measure, and the factors estimates are the ones associated with this optimal decay.

We implemented this approach and report the estimated error measure in Figure C.1. The optimal decay parameter is 0.0435. The value attained by the loss function at $\lambda = 0.0609$ is similar, however. One major disadvantage of this two-step approach is that it introduces a look-ahead bias: the final estimator of the factors at time t will depend on the dynamics of yields at dates after t since the loss function incorporates information from the whole sample by construction. Hence, an econometrician who follows this method and is furnished with only a truncated version of our data could find different results. Again, due to our focus on forecasting, we prefer to pay the cost of a slightly worse in-sample fit to get factor estimates that do not contain any look-ahead bias.

Figure C.2 assesses how high the price is in terms of in-sample fit that we are paying when using the same strategy as Diebold and Li (2006). It is not a high price. For each date, we compute the average squared residual in the cross-section of yields after fitting the model, take the square root, and plot it as a function of time. This time series is a direct measure of how much information we lose from the yield curve by using a reduced-form model to summarize it. Until 2009, all three time series are indistinguishable. Between 2010 and 2014, the performance of the Diebold and Li (2006) approach deteriorated with respect to the other methods but not by a large amount. After 2014 all methods seem to generate equally reasonable fits.

A third alternative approach is the one in Diebold et al. (2006). They leverage the linearity of (4) to estimate factors with a Kalman-filter in which the state equation also has macroeconomic variables. Their factor estimates are the Kalman-smoothed series based on parameters estimated by maximum likelihood. Their focus is on the joint dynamics of yields and macroeconomic variables and they do not emphasize forecasting. It is not obvious to us that a state-space representation would improve in-

Figure C.1: Profiling of the decay parameter. For each value of λ, we fit the Nelson-Siegel model by OLS date by date. Then we compute a monthly measure of the average squared fitting error in the cross-section. We finally average over time and plot this information denoted by "Fit Error" as a function of λ. The black dashed line represents the overall argmin while the red dashed line is the value used by Diebold and Li (2006). The sample size ranges from January 1973 to December 2021. We use information on all yields from up to 120 months.



sample fit, however. [38] Moreover, their system, although small, has 36 parameters to be estimated. Using Kalman-smoothing at every point in time would imply a new 36-dimensional numerical optimization for each date, which would likely create the same type of problems as the NLS approach. Alternatively, Kalman-filtered estimates of the factors at the beginning of the sample would likely be too dependent on the imposed priors due to the low number of data points, going against our goal of fitting the yield curve in the best way we can.

Finally, Figure C.3 shows how our parametrization of the Nelson-Siegel model fares against a polynomial model of the form:

$$y_t^{(\tau)} = c_0 + c_1 \times \tau + c_2 \times \tau^2, \tag{C.1}$$

where c_0, c_1, and c_2 are constants. Although these models have the same number of free parameters, the Nelson-Siegel representation achieves a better fit for the vast majority of dates. It was only surpassed by the polynomial model during the heights of the Global Financial crisis in 2008 and a brief period between 2012 and 2014.

[38] See the discussion on Chapter 1 of Diebold and Rudebusch (2013).

Figure C.2: Time-series of the average squared residual when fitting the cross-section of yields using different methods. "Diebold-Li" corresponds to OLS with $\lambda = 0.609$. "NLS" represents the error attained when we estimated models using non-linear least squares date by date. "Optimal OLS" uses the OLS approach with $\lambda = 0.0435$. The sample ranges from 1973 to 2021.

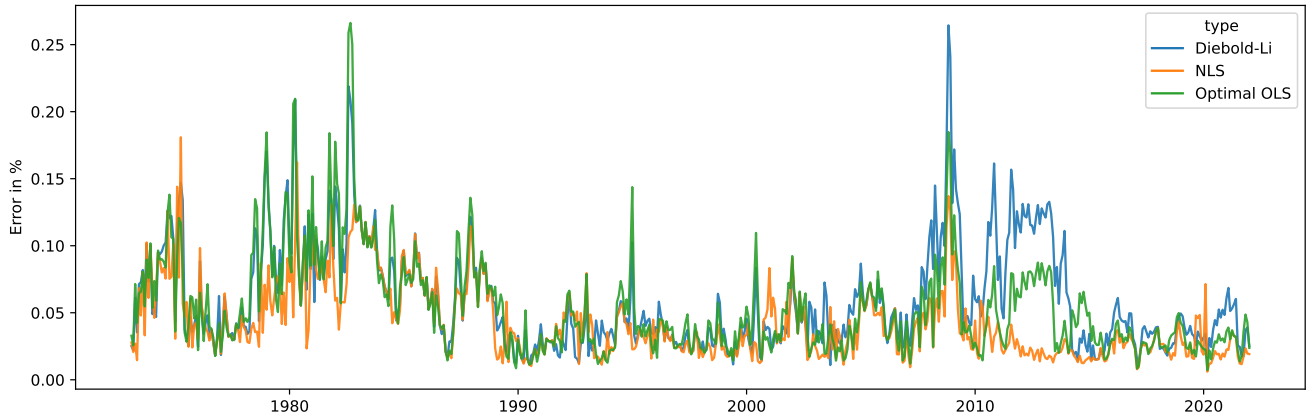
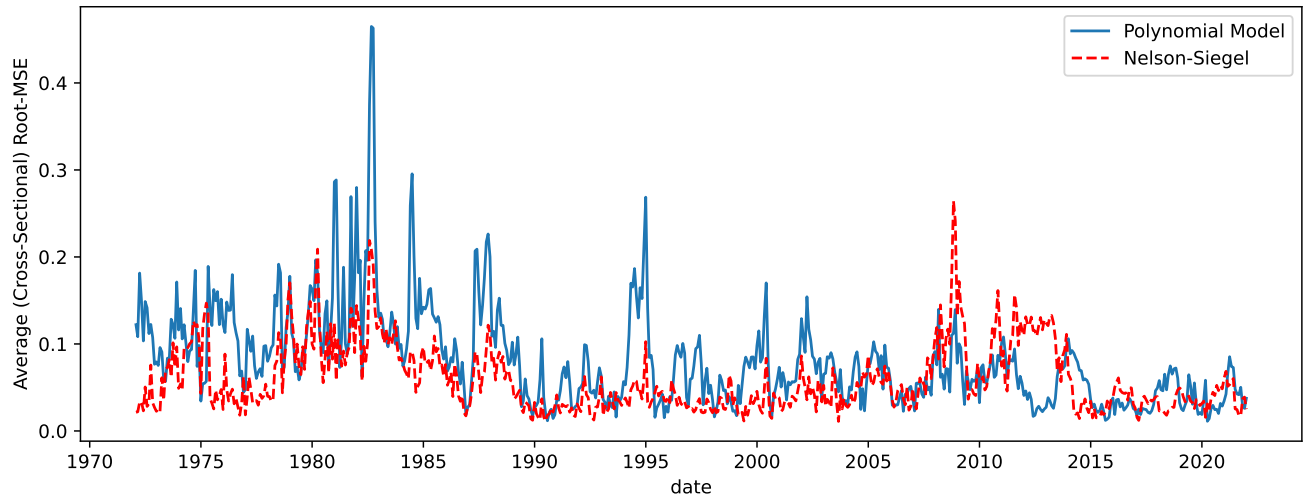


Figure C.3: This figure reports the same measure as Figure C.2 but compares our Nelson-Siegel approach with the reduced-form model from equation (C.1). Both estimation procedures use all available yields from 1 to 120 months for each date.



D FRED-MD Variables

In this appendix, we report the full set of variables from the FRED-MD data set we use. Table D.1 has four columns. The first represents the FRED code for each respective series. The second one lists the category of each variable as described in McCracken and Ng (2016). The third is a simple description of each series. The fourth encodes what transformation we used to make each series stationary. For each series x_t, we denote the transformed series as z_t. We follow the convention:

- Code 1: $z_t = x_t$
- Code 2: $z_t = x_t - x_{t-1}$
- Code 3: $z_t = x_t - x_{t-2}$
- Code 4: $z_t = \log(x_t)$
- Code 5: $z_t = \log(x_t / x_{t-1})$
- Code 6: $z_t = \log(x_t / x_{t-2})$
- Code 7: $z_t = \frac{x_t - x_{t-1}}{x_{t-1}}$

Table D.1: Full list of our macroeconomic variables

FRED Code	Category	Description	Transformation Code
HOUST	Housing	Housing Starts: Total New Privately Owned	4
HOUSTMW	Housing	Housing Starts, Midwest	4
HOUSTNE	Housing	Housing Starts, Northeast	4
HOUSTS	Housing	Housing Starts, South	4
HOUSTW	Housing	Housing Starts, West	4
PERMIT	Housing	New Private Housing Permits (SAAR)	4
PERMITMW	Housing	New Private Housing Permits, Midwest (SAAR)	4
PERMITNE	Housing	New Private Housing Permits, Northeast (SAAR)	4
PERMITS	Housing	New Private Housing Permits, South (SAAR)	4
PERMITW	Housing	New Private Housing Permits, West (SAAR)	4
AAA	Interest and Exchange Rates	Moody's Seasoned Aaa Corporate Bond Yield	2
AAAFFM	Interest and Exchange Rates	Moody's Aaa Corporate Bond Minus FEDFUNDS	1
BAA	Interest and Exchange Rates	Moody's Seasoned Baa Corporate Bond Yield	2
BAAFFM	Interest and Exchange Rates	Moody's Baa Corporate Bond Minus FEDFUNDS	1
COMPAPFFx	Interest and Exchange Rates	3-Month Commercial Paper Minus FEDFUNDS	1
CP3Mx	Interest and Exchange Rates	3-Month AA Financial Commercial Paper Rate	2
EXCAUSx	Interest and Exchange Rates	Canada / U.S. Foreign Exchange Rate	5
EXJPUSx	Interest and Exchange Rates	Japan / U.S. Foreign Exchange Rate	5
EXSZUSx	Interest and Exchange Rates	Switzerland / U.S. Foreign Exchange Rate	5
EXUSUKx	Interest and Exchange Rates	U.S. / U.K. Foreign Exchange Rate	5
FEDFUNDS	Interest and Exchange Rates	Effective Federal Funds Rate	2
GS1	Interest and Exchange Rates	1-Year Treasury Rate	2
GS10	Interest and Exchange Rates	10-Year Treasury Rate	2
GS5	Interest and Exchange Rates	5-Year Treasury Rate	2
T10YFFM	Interest and Exchange Rates	10-Year Treasury C Minus FEDFUNDS	1
T1YFFM	Interest and Exchange Rates	1-Year Treasury C Minus FEDFUNDS	1
T5YFFM	Interest and Exchange Rates	5-Year Treasury C Minus FEDFUNDS	1
TB3MS	Interest and Exchange Rates	3-Month Treasury Bill:	2
TB3SMFFM	Interest and Exchange Rates	3-Month Treasury C Minus FEDFUNDS	1
TB6MS	Interest and Exchange Rates	6-Month Treasury Bill:	2
TB6SMFFM	Interest and Exchange Rates	6-Month Treasury C Minus FEDFUNDS	1
TWEXAFEGSMTHx	Interest and Exchange Rates	Trade Weighted U.S. Dollar Index	5

Continued on next page

FRED Code	Category	Description	Transformation Code
AWHMAN	Labor Market	Avg Weekly Hours : Manufacturing	1
AWOTMAN	Labor Market	Avg Weekly Overtime Hours : Manufacturing	2
CE16OV	Labor Market	Civilian Employment	5
CES0600000007	Labor Market	Avg Weekly Hours : Goods-Producing	1
CES0600000008	Labor Market	Avg Hourly Earnings : Goods-Producing	6
CES1021000001	Labor Market	All Employees: Mining and Logging: Mining	5
CES2000000008	Labor Market	Avg Hourly Earnings : Construction	6
CES3000000008	Labor Market	Avg Hourly Earnings : Manufacturing	6
CLAIMSx	Labor Market	Initial Claims	5
CLF16OV	Labor Market	Civilian Labor Force	5
DMANEMP	Labor Market	All Employees: Durable goods	5
HWI	Labor Market	Help-Wanted Index for United States	2
HWIURATIO	Labor Market	Ratio of Help Wanted/No. Unemployed	2
MANEMP	Labor Market	All Employees: Manufacturing	5
NDMANEMP	Labor Market	All Employees: Nondurable goods	5
PAYEMS	Labor Market	All Employees: Total nonfarm	5
SRVPRD	Labor Market	All Employees: Service-Providing Industries	5
UEMP15OV	Labor Market	Civilians Unemployed - 15 Weeks \& Over	5
UEMP15T26	Labor Market	Civilians Unemployed for 15-26 Weeks	5
UEMP27OV	Labor Market	Civilians Unemployed for 27 Weeks and Over	5
UEMP5TO14	Labor Market	Civilians Unemployed for 5-14 Weeks	5
UEMPLT5	Labor Market	Civilians Unemployed - Less Than 5 Weeks	5
UEMPMEAN	Labor Market	Average Duration of Unemployment (Weeks)	2
UNRATE	Labor Market	Civilian Unemployment Rate	2
USCONS	Labor Market	All Employees: Construction	5
USFIRE	Labor Market	All Employees: Financial Activities	5
USGOOD	Labor Market	All Employees: Goods-Producing Industries	5
USGOVT	Labor Market	All Employees: Government	5
USTPU	Labor Market	All Employees: Trade, Transportation \& Utilities	5
USTRADE	Labor Market	All Employees: Retail Trade	5
USWTRADE	Labor Market	All Employees: Wholesale Trade	5
BOGMBASE	Money and Credit	Monetary Base	6
BUSLOANS	Money and Credit	Commercial and Industrial Loans	6
CONSPI	Money and Credit	Nonrevolving consumer credit to Personal Income	2
DTCOLNVHFNM	Money and Credit	Consumer Motor Vehicle Loans Outstanding	6
DTCTHFNM	Money and Credit	Total Consumer Loans and Leases Outstanding	6
INVEST	Money and Credit	Securities in Bank Credit at All Commercial Banks	6
M1SL	Money and Credit	M1 Money Stock	6
M2REAL	Money and Credit	Real M2 Money Stock	5
M2SL	Money and Credit	M2 Money Stock	6
NONBORRES	Money and Credit	Reserves Of Depository Institutions	7
NONREVSL	Money and Credit	Total Nonrevolving Credit	6
REALLN	Money and Credit	Real Estate Loans at All Commercial Banks	6
TOTRESNS	Money and Credit	Total Reserves of Depository Institutions	6
ACOGNO	Orders and Inventories	New Orders for Consumer Goods	5
AMDMNOx	Orders and Inventories	New Orders for Durable Goods	5
AMDMUOx	Orders and Inventories	Unfilled Orders for Durable Goods	5
ANDENOx	Orders and Inventories	New Orders for Nondefense Capital Goods	5
BUSINVx	Orders and Inventories	Total Business Inventories	5
CMRMTSPLx	Orders and Inventories	Real Manu. and Trade Industries Sales	5
DPCERA3M086SBEA	Orders and Inventories	Real personal consumption expenditures	5
ISRATIOx	Orders and Inventories	Total Business: Inventories to Sales Ratio	2
RETAILx	Orders and Inventories	Retail and Food Services Sales	5
UMCSENTx	Orders and Inventories	Consumer Sentiment Index	2
CUMFNS	Output and Income	Capacity Utilization: Manufacturing	2
INDPRO	Output and Income	IP Index	5
IPBUSEQ	Output and Income	IP: Business Equipment	5
IPCONGD	Output and Income	IP: Consumer Goods	5
IPDCONGD	Output and Income	IP: Durable Consumer Goods	5
IPDMAT	Output and Income	IP: Durable Materials	5
IPFINAL	Output and Income	IP: Final Products (Market Group)	5
IPFPNSS	Output and Income	IP: Final Products and Nonindustrial Supplies	5
IPFUELS	Output and Income	IP: Fuels	5
IPMANSICS	Output and Income	IP: Manufacturing (SIC)	5
IPMAT	Output and Income	IP: Materials	5
IPNCONGD	Output and Income	IP: Nondurable Consumer Goods	5
IPNMAT	Output and Income	IP: Nondurable Materials	5

Continued on next page

FRED Code	Category	Description	Transformation Code
RPI	Output and Income	Real Personal Income	5
W875RX1	Output and Income	Real personal income ex transfer receipts	5
CPIAPPSL	Prices	CPI : Apparel	6
CPIAUCSL	Prices	CPI : All Items	6
CPIMEDSL	Prices	CPI : Medical Care	6
CPITRNSL	Prices	CPI : Transportation	6
CPIULFSL	Prices	CPI : All Items Less Food	6
CUSR0000SA0L2	Prices	CPI : All items less shelter	6
CUSR0000SA0L5	Prices	CPI : All items less medical care	6
CUSR0000SAC	Prices	CPI : Commodities	6
CUSR0000SAD	Prices	CPI : Durables	6
CUSR0000SAS	Prices	CPI : Services	6
DDURRG3M086SBEA	Prices	Personal Cons. Exp: Durable goods	6
DNDGRG3M086SBEA	Prices	Personal Cons. Exp: Nondurable goods	6
DSERRG3M086SBEA	Prices	Personal Cons. Exp: Services	6
OILPRICEx	Prices	Crude Oil, spliced WTI and Cushing	6
PCEPI	Prices	Personal Cons. Expend.: Chain Index	6
PPICMM	Prices	PPI: Metals and metal products:	6
WPSFD49207	Prices	PPI: Finished Goods	6
WPSFD49502	Prices	PPI: Finished Consumer Goods	6
WPSID61	Prices	PPI: Intermediate Materials	6
WPSID62	Prices	PPI: Crude Materials	6
S&P 500	Stock Market	S\&P's Common Stock Price Index: Composite	5
S&P PE ratio	Stock Market	S\&P's Composite Common Stock: Price-Earnings Ratio	5
S&P div yield	Stock Market	S\&P's Composite Common Stock: Dividend Yield	2
S&P: indust	Stock Market	S\&P's Common Stock Price Index: Industrials	5
VIXCLSx	Stock Market	VIX	1

E Proofs

E.1 Proof of Proposition 1

We start repeating the equation for yields as in (4) with a constant positive decay parameter:

$$y_t^{(\tau)} = \beta_{1,t} + \beta_{2,t}\left(\frac{1-e^{-\lambda\tau}}{\lambda\tau}\right) + \beta_{3,t}\left(\frac{1-e^{-\lambda\tau}}{\lambda\tau} - e^{-\lambda\tau}\right) \tag{E.1}$$

It's crucial to notice that τ is measured in months. To avoid abusing notation, denote by $\xi_t(n)$ the zero-coupon yield at time t for a maturity of n years. For a fixed n, there pick $\tau = 12 \cdot n$. Naturally, it follows that:

$$\xi_t(n) = y_t^{(12\cdot n)} = \beta_{1,t} + \beta_{2,t}\left(\frac{1-e^{-12\lambda\tau}}{12\lambda\tau}\right) + \beta_{3,t}\left(\frac{1-e^{-12\lambda\tau}}{\lambda 12\tau} - e^{-12\lambda\tau}\right)$$

$$\Downarrow$$

$$\xi_t(n) = \beta_{1,t} + \beta_{2,t}\left(\frac{1-e^{-\theta n}}{\theta n}\right) + \beta_{3,t}\left(\frac{1-e^{-\theta n}}{\theta n} - e^{-\theta n}\right)$$

where $\theta = 12\lambda > 0$. Multiplying both sides by n yields

$$n \cdot \xi_t(n) = n\beta_{1,t} + \left(\frac{1-e^{-\theta n}}{\theta}\right)[\beta_{2,t} + \beta_{3,t}] - n\beta_{3,t}e^{-\theta n} \tag{E.2}$$

$$(n-1) \cdot \xi_{t+12}(n-1) = (n-1)\beta_{1,t+12} + \left(\frac{1-e^{-\theta(n-1)}}{\theta}\right)[\beta_{2,t+12} + \beta_{3,t+12}] - (n-1)\beta_{3,t+12}e^{-\theta(n-1)} \tag{E.3}$$

$$\xi_t(1) = \beta_{1,t} + \left(\frac{1-e^{-\theta}}{\theta}\right)[\beta_{2,t} + \beta_{3,t}] - \beta_{3,t}e^{-\theta} \tag{E.4}$$

To compute the excess bond returns as in (1), we need to subtract the last two equations from the first one. We keep track of the three terms that appear in each of the equations above.

First term. Collecting the terms in β_1 yields

$$n\beta_{1,t} - (n-1)\beta_{1,t+12} - \beta_{1,t} = (n-1)\left[\beta_{1,t} - \beta_{1,t+12}\right] \tag{E.5}$$

Second term. We now collect the terms in $[\beta_{2,t} + \beta_{3,t}]$

$$\left(\frac{1-e^{-\theta n}}{\theta} - \frac{1-e^{-\theta}}{\theta}\right)[\beta_{2,t} + \beta_{3,t}] = \left(\frac{1-e^{-\theta(n-1)}}{\theta}\right)e^{-\theta}[\beta_{2,t} + \beta_{3,t}]$$

The constant inside the parenthesis above is the same that appears in the analogous term for the expression of $(n-1) \cdot \xi_{t+12}(n-1)$. Hence, we have

$$\left(\frac{1-e^{-\theta n}}{\theta} - \frac{1-e^{-\theta}}{\theta}\right)[\beta_{2,t} + \beta_{3,t}] - \left(\frac{1-e^{-\theta(n-1)}}{\theta}\right)[\beta_{2,t+12} + \beta_{3,t+12}] = \left(\frac{1-e^{-\theta(n-1)}}{\theta}\right)\left[\left(e^{-\theta}\beta_{2,t} - \beta_{2,t+12}\right) + \left(e^{-\theta}\beta_{3,t} - \beta_{3,t+12}\right)\right] \tag{E.6}$$

Third term. Here we have

$$\begin{aligned}
-n\beta_{3,t}e^{-\theta n} + (n-1)\beta_{3,t+12}e^{-\theta(n-1)} + \beta_{3,t}e^{-\theta} &= -n\beta_{3,t}e^{-\theta n} + (n-1)\beta_{3,t+12}e^{-\theta(n-1)} + \beta_{3,t}e^{-\theta} + \beta_{3,t+12} - \beta_{3,t+12} \\
&= \left(ne^{-\theta(n-1)} - 1\right)\left(\beta_{3,t+12} - e^{-\theta}\beta_{3,t}\right) + \beta_{3,t+12}\left(1 - e^{\theta(n-1)}\right) \\
&= \left(1 - ne^{-\theta(n-1)}\right)\left(e^{-\theta}\beta_{3,t} - \beta_{3,t+12}\right) + \beta_{3,t+12}\left(1 - e^{\theta(n-1)}\right)
\end{aligned}$$

Then, the proposition follows from summing the expressions derived for the first, second, and third terms.

E.2 Expected Short Rates and Term Premia

We now prove the decomposition in (14). It follows standard steps.

Recall the definition of the excess returns:

$$xr_{t+12}^{(n)} = n \cdot y_t^{(n)} - (n-1) \cdot y_{t+12}^{(n-1)} - y_t^{(1)}$$

We can keep iterating indexes 1 year ahead and have

$$\begin{aligned}
xr_{t+24}^{(n-1)} &= (n-1) \cdot y_{t+12}^{(n-1)} - (n-2) \cdot y_{t+24}^{(n-2)} - y_{t+12}^{(1)} \\
xr_{t+36}^{(n-2)} &= (n-2) \cdot y_{t+24}^{(n-2)} - (n-3) \cdot y_{t+36}^{(n-3)} - y_{t+24}^{(1)} \\
&\vdots \\
xr_{t+12\cdot k}^{(n-k+1)} &= (n-k+1) \cdot y_{t+12\cdot(k-1)}^{(n-k+1)} - (n-k) \cdot y_{t+36}^{(n-k)} - y_{t+12\cdot(k-1)}^{(1)} \\
&\vdots \\
xr_{t+12\cdot n}^{(1)} &= 0
\end{aligned}$$

If we sum up these equations and keep track of the cancellations, we have

$$\sum_{k=1}^{n} xr_{t+12\cdot k}^{(n-k+1)} = n \cdot y_t^{(n)} - \sum_{k=1}^{n} y_{t+12\cdot(k-1)}^{(1)} \tag{E.7}$$

If we apply the conditional expectation operator $\mathbb{E}_t$ on both sides, we get an expression that is equivalent to (14).

We also note that if we were working with monthly returns, we would have the same type of expression. The only difference is that the short rate would be the monthly short rate. The specific definition, either in terms of annual returns or in terms of monthly returns, does not change the interpretation of each of these components. Our notation is different than the one from Adrian et al. (2013), for example, because we focus on yearly returns.

F Conditional Predictive Ability

F.1 Setup

The tests we have employed in Section 5 are designed to test whether, on average, a model with more state variables is able to forecast better out of the sample than a model that only uses information from the yield curve. Now we are interested in investigating *when* these violations are happening.

Let us define the loss-function when forecasting as $L_{i,t} \equiv \left(\beta_{i,t} - \widehat{\beta}_{i,t}\right)^2$, where $\widehat{\beta}_{i,t}$ is the forecast value for a factor i. Moreover, we define $D_{i,t}$ as the following difference of loss-functions:

$$D_{i,t} \equiv L_{i,t}^{(\text{SH})} - L_{i,t}^{(\text{Macro})}, \tag{F.1}$$

where the first term is the loss function attained under the Spanning Hypothesis and the second term is the loss function attained when we use extra state variables. A *positive value* for $D_{i,t}$ implies that *macro data was helpful* for forecasting. The analysis so far tested the following hypothesis:

$$H_0 : \mathbb{E}[D_{i,t}] = 0,$$

which is a moment condition implied by the spanning hypothesis.

We now seek to test another null hypothesis, that is, a conditional version of the moment condition above. Given a certain filtration $\{\mathcal{G}_t\}$, we are interested in testing

$$H_0 : \mathbb{E}\left[D_{i,t+12}|\mathcal{G}_t\right] = 0. \tag{F.2}$$

We notice that this conditional expectation encompasses the previous case since one can always take the trivial sigma-algebra for $\mathcal{G}_t$. Under this more general null hypothesis, however, the violations of the Spanning Hypothesis cannot be predicted by any process x_t that is $\mathcal{G}_t$-adapted. In other words, under the null, an econometrician with information up to time t would not be able to tell whether that is a good or a bad moment to use either model.

The theory developed in Giacomini and White (2006) fits our needs. They develop a test statistic and an asymptotic approximation that can deal with conditional moments like in equation (F.2). It consists of a Wald-type test that is easy to implement. We specialize their notation to our setup. Let $\mathbf{x}_t$ be a $q \times 1$ random vector with variables chosen by the econometrician and let $\{\mathcal{G}_t\}_{t=t_0}^{T-h}$ be the natural filtration of $\mathbf{x}_t$, where h is a generic forecasting horizon. Let $\mathbf{z}_{t+h} \equiv \mathbf{x}_t D_{t+h}$ and let:

$$\overline{\mathbf{z}}_T \equiv \frac{1}{T-h-t_0}\sum_{t=t_0}^{T-h}\mathbf{z}_{t+h},$$

$$\widehat{\Omega}_T \equiv \frac{1}{T-h-t_0}\sum_{t=t_0}^{T-h}\mathbf{z}_{t+h}\mathbf{z}_{t+h}' + \frac{1}{T-h-t_0}\sum_{j=1}^{h-1}w_{j,T}\sum_{t=t_0+j}^{T-h}\left(\mathbf{z}_{t+h-j}\mathbf{z}_{t+h}' + \mathbf{z}_{t+h}\mathbf{z}_{t+h-j}'\right).$$

The first definition is just the average of $\mathbf{z}_t$ over time while $\widehat{\Omega}_T$ is a HAC-type long-run variance estimator, in which it is assumed that $w_{j,T} \to 1$ as $T \to \infty$ for each $j \in \{1, ..., h-1\}$. Under some regularity conditions, they show that:

$$W \equiv T \cdot \mathbf{z}_{t+h}'\widehat{\Omega}_T^{-1}\mathbf{z}_{t+h} \xrightarrow{d} \chi_q^2. \tag{F.3}$$

In our setup, we take $h = 12$. If $h = 1$, this is equivalent to testing whether D_t is a martingale difference sequence, for instance. In case one rejects the null hypothesis, it is natural to inspect the projection of D_{t+12} on $\mathbf{x}_t$, which we do below.

One drawback of this methodology, as emphasized by Giacomini and White (2006), is that it can only be applied to *rolling window* estimation schemes. Hence, we use the rolling window Random Forest forecasts used in section 7.

There is a trade-off when picking the window size. Given the same amount of data, a small window allows for forecasts with lower correlation over time, but also implies that each forecast is done with less data and, hence, with less precision. A larger window implies that each individual forecast is done using more data but forecasts will have higher autocorrelation. We focus on results using a 180-month window to keep it consistent with the main text. We keep the same out-of-sample period from before (1990-2021) and forecast the level of the factors as we did before.

Figure F.1: The figure shows time-series for $D_{i,t}$ estimated using a Random Forest with 500 trees. The out-of-sample period is 1990-2021. Here we control for forward rates and use a 180-month rolling window for estimation. All three series are scaled by their respective standard deviation so the scales are comparable. A positive value means that macro data was useful when forecasting the Nelson-Siegel factors.

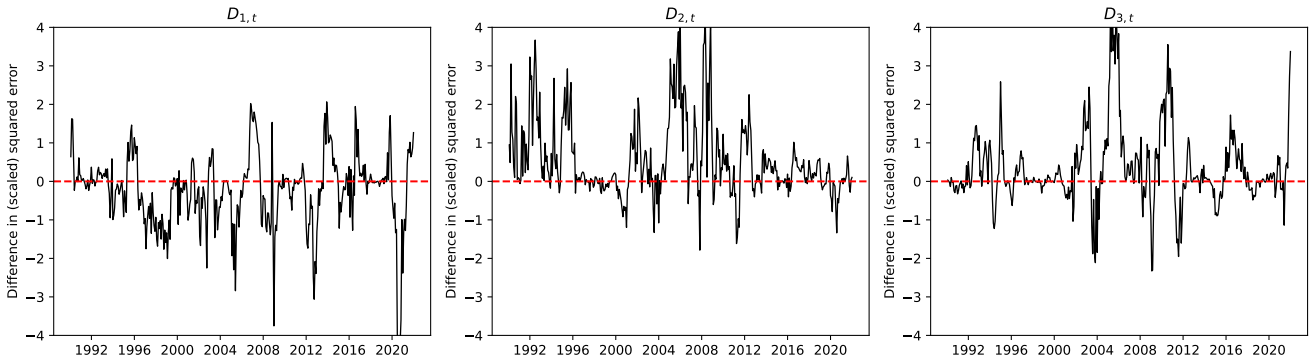


Figure F.2: This figure shows time-series for $D_{i,t}$ estimated using a Random Forest with 500 trees. The out-of-sample period is 1990-2021. Here we control for the lagged Nelson-Siegel factors and use a 180-month rolling window for estimation. All three series are scaled by their respective standard deviation, so the scales are comparable. A positive value means that macro data was useful when forecasting the Nelson-Siegel factors.

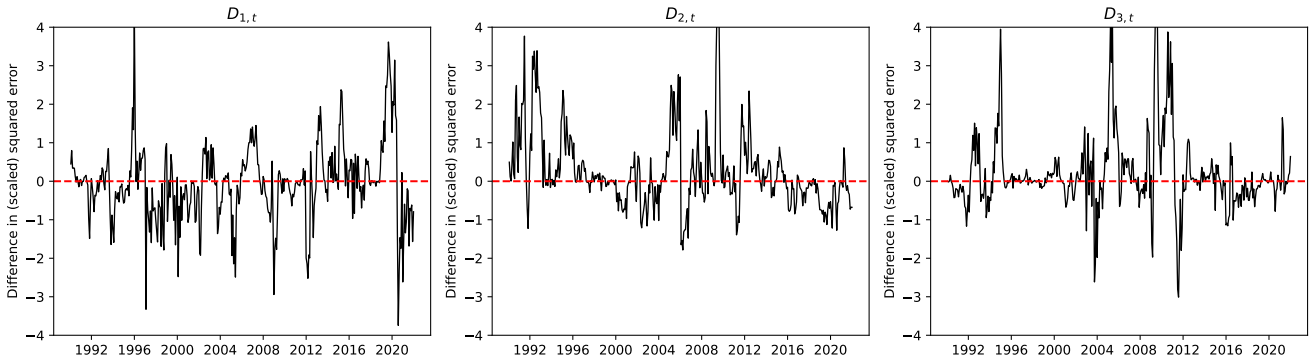


Figure F.1 reports the realizations of $D_{i,t}$ when controlling for the forward rates, while Figure F.2 displays the difference in loss functions when controlling by the lagged Nelson-Siegel factors. We

scale $D_{i,t}$ by their respective standard deviation such that scales are comparable. Positive realizations imply that using macroeconomic variables for forecasting was useful, while negative realizations imply otherwise. In both cases, we see that D_1 lives mostly on the negative plane, while D_2 is situated mostly on the positive plane. This is consistent with our previous evidence that macroeconomic data was helpful for forecasting β_2, but not β_1. Table 6, in the main text, formally tests and confirms this observation.

F.2 Running the Test

We now turn to a full-fledged conditional predictive ability test. An important choice upon the econometrician is the choice of $\mathcal{G}_t$. In fact, the standard setup from DTSMs provides no guidance. After all, violations of the Spanning Hypothesis should not be there anyway, let alone be predicted by other variables. We pick five variables for $\mathcal{G}_t$ with the following motivation in mind:

1. The Economic Policy Uncertainty index (EPU) from Baker et al. (2016): this is a news-based index that measures the frequency of coverage regarding economic policy by the major US newspapers. Borup et al. (2023) does a similar analysis to ours, although in the context of the *expectation* hypothesis, and finds stronger predictability using macroeconomic variables when there is less economic policy uncertainty. We use a 3-month rolling window mean of the raw EPU index.

2. The Chicago Fed National Activity Index (CFNAI): this is a monthly measure computed by the Chicago Fed that tries to summarize the business cycle for the U.S. economy. Borup et al. (2023) find that bond excess returns are better predicted in times of higher economic activity. We also use a 3-month rolling window mean.

3. The unemployment gap (UGAP): it is defined as the difference between the unemployment rate for the U.S. and its natural unemployment rate, both reported by the St. Louis Fed. This is natural since we use unemployment in our Taylor Rule specification in the main text.

4. PCE inflation: core PCE inflation measured taken from the St Louis Fed. Our motivation here is the same as UGAP since inflation control is one of the elements we use in our Taylor Rule. Additionally, as emphasized by Bauer and Rudebusch (2017), UGAP and PCE are usually included as risk factors in some traditional DTSMs.

5. Slope: we define this as the difference between the 10-year rate and the Fed Funds rate at the end of each month, both taken from the St Louis Fed as well. This is a measure of the slope of the U.S. yield curve and an empirical proxy for β_2. We use this to asses whether the slope of the yield curve has any predictive power for the violations of the spanning hypothesis since there is a long literature showing how the slope of the yield curve can predict business-cycle movements. See Hännikäinen (2017) and the references therein, for example.

The variables used in this analysis are plotted in Figure A.12 in Appendix A. Since we found no violation of the Spanning Hypothesis neither through β_1 nor through β_3 on average, we only study the conditional predictive ability test for β_2.

Table F.2: The last row reports the p-value associated to the test from Giacomini and White (2006) for D_2, computed controlling for the forward rates and displayed in Figure F.1. Each column also reports the estimate for $\boldsymbol{b}$ from (F.4). Stars denote significance at 10%, 5%, and 1%, respectively. Standard errors for $\boldsymbol{b}$ are computed using Newey and West (1987). The out-of-sample period is 1990-2021.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
EPU	-0.08					-0.11				-0.19*
	(0.10)					(0.10)				(0.10)
CFNAI		-0.06				-0.10		-0.09		-0.14
		(0.08)				(0.07)		(0.09)		(0.08)
UGAP			-0.02				0.04	0.01	-0.03	0.05
			(0.10)				(0.09)	(0.09)	(0.08)	(0.13)
PCE				0.30**			0.31**	0.31**	0.30**	0.33***
				(0.12)			(0.12)	(0.12)	(0.12)	(0.11)
Slope					0.09				0.12	0.10
					(0.12)				(0.11)	(0.12)
N	384	384	384	384	384	384	384	384	384	384
R2	0.01	0.00	0.00	0.09	0.01	0.02	0.09	0.10	0.10	0.13
GW p-values	0.51	0.38	0.84	0.00	0.45	0.50	0.00	0.00	0.01	0.01

Table F.2 reports at its last row the p-values for the conditional predictive ability test based on (F.3). The different columns study different combinations of the variables above to create $\mathcal{G}_t$. Small p-values imply that we can reject the null hypothesis, i.e., we can predict when macroeconomic variables will be the most useful and violate the Spanning Hypothesis.

The only specifications for which we can reject the null are the ones that include PCE. This suggests that, across the variables we focused on, only the 12-month inflation has predictive power for D_t. In order to closely inspect this correlation, we also report coefficient estimates for $\mathbf{b}$ from the following linear regression:

$$D_{2,t+12} = a + \mathbf{b}'\mathbf{x}_t + u_{t+12}, \tag{F.4}$$

in which we standardize both the dependent and independent variables so we can interpret $\mathbf{b}$ in terms of standard deviations. For each specification of $\mathcal{G}_t$ (which is always the natural filtration of $\mathbf{x}_t$), we compute the standard errors associated to the estimation of $\mathbf{b}$ using the HAC estimator from Newey and West (1987).

Columns 1 through 5 analyze the variables individually. We can only reject the null when we consider the 12-month inflation and the estimated coefficient in the linear projection is statistically significant at the 95% confidence value. It suggests that when inflation, at time t, was one standard deviation above its mean, the difference in loss functions denoted by $D_{2,t+12}$ was 0.3 standard deviations above its mean. This implies that periods of higher inflation are associated with moments where the gain in using macroeconomic variables to forecast ahead was higher.

Column 6 uses the variables used in Borup et al. (2023). However, with a p-value of 0.50, we cannot reject the null hypothesis at the usual levels. We stress that our results are not necessarily in disagreement with theirs since we study deviations from a baseline model that allows risk premia to

Table F.3: The last row reports the p-value associated to the test from Giacomini and White (2006) for D_2, computed controlling for the forward rates and displayed in Figure F.2. Each column also reports the estimate for $\boldsymbol{b}$ from (F.4). Stars denote significance at 10%, 5%, and 1%, respectively. Standard errors for $\boldsymbol{b}$ are computed using Newey and West (1987). The out-of-sample period is 1990-2021.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
EPU	0.05					0.03				-0.16
	(0.11)					(0.11)				(0.10)
CFNAI		-0.09				-0.08		-0.04		-0.09
		(0.08)				(0.08)		(0.07)		(0.07)
UGAP			0.23*				0.27**	0.26**	0.17	0.25*
			(0.13)				(0.12)	(0.12)	(0.12)	(0.15)
PCE				0.19			0.24*	0.24*	0.23*	0.26**
				(0.14)			(0.13)	(0.13)	(0.12)	(0.12)
Slope					0.26**				0.18	0.15
					(0.13)				(0.12)	(0.12)
N	384	384	384	384	384	384	384	384	384	384
R^2	0.00	0.01	0.05	0.04	0.07	0.01	0.11	0.11	0.13	0.15
GW p-values	0.56	0.22	0.07	0.19	0.05	0.39	0.01	0.02	0.04	0.04

change over time, unlike their setup that assumes a constant risk premia. Here, following Proposition (1), we are interested in understanding whether movements in the Nelson-Siegel factors can be better anticipated by macroeconomic state variables and not whether they move at all.

Column 7 studies a specification with PCE and UGAP. Controlling for the unemployment rate did not change our evidence regarding PCE, both qualitatively and quantitatively. Columns 8 and 9 add CFNAI and Slope, respectively. The coefficient on PCE remains unchanged, although no other variable displays a significant coefficient.

The final column uses all five variables. In that regression, the coefficient on PCE is now significant even at the 99% confidence level, with virtually the same magnitude as before. The coefficient on EPU is negative and significant, now in line with the findings in Borup et al. (2023).

The general evidence from Table F.2 is simple: our augmented model with macroeconomic variables is useful for forecasting β_2 not only on average but especially when inflation at time t is higher. The estimated coefficient of 0.3 is stable across specifications. No other variable studied here has the same success in predicting $D_{2,t+12}$.

Table F.3 repeat this exercise when using lagged Nelson-Siegel factors to control for the information already in the yield curve. It tells a similar story, although the estimated coefficient is around 0.25. In summary, no matter how we control for the yield curve, periods of higher inflation seem to be moments in which our machinery for predicting β_2 is the most useful. Violations of the Spanning Hypothesis at the short end of the yield curve tend to occur when inflation is high.

F.3 Non-Parametric Evidence

Although the theory developed in Giacomini and White (2006) fits elegantly in our framework, we also provide evidence regarding the correlation between the 12-month inflation and D_t using a simple non-parametric approximation. We define the different months of our out-of-sample period into inflation terciles: low inflation, medium inflation, and high inflation. The average annualized inflation within each inflation tercile was 1.3%, 1.8%, and 2.8%, respectively.

We compute the average value for $D_{i,t}, i = 1, 2, 3$ for each tercile. The results are reported in Table F.4. This is a non-parametric estimator for the conditional expectation of $D_{i,t+12}$ given the inflation tercile at time t. The top three rows show results controlling for the forward rates, while the bottom three rows show results controlling for the lagged Nelson-Siegel factors. In both cases, we use a 180-month rolling window for estimation.

Table F.4: We create inflation terciles based on the PCE measure. For each tercile, we take the average of the difference in loss functions D_i, computed using a Random Forest with a 180-month rolling window and 500 trees. The top three rows control for the forward rates while the bottom three rows control for the lagged Nelson-Siegel factors. The out-of-sample period is 1990-2021.

Inflation Tercile	PCE	D_1	D_2	D_3	Control
Low	0.013	-0.152	0.496	2.386	Forward Rates
Medium	0.018	-0.754	0.788	1.923	Forward Rates
High	0.028	0.039	2.430	1.526	Forward Rates
Low	0.013	-0.204	-0.023	0.803	Lagged Factors
Medium	0.018	-0.114	0.120	0.850	Lagged Factors
High	0.028	0.048	1.963	1.492	Lagged Factors

The average value for D_2 is increasing with the inflation terciles, no matter how we control for the yield curve. It is noticeably higher at the highest tercile. However, we see no apparent pattern for D_1 and D_3. The message suggested by Table F.4 is the same as the one we took from the more formal analysis of Table F.2: moments of higher inflation at time t are associated with more likely violations of the Spanning Hypothesis through β_2 in our sample.