Asymmetric Violations of the Spanning Hypothesis*

Gustavo Freire†1 and **Raul Riva**‡2

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

First Version: March 17, 2023
This Version: December 31, 2023

Abstract

We document that the Spanning Hypothesis, which is implied by most macro-finance term structure models, is violated asymmetrically along the U.S. yield curve. After controlling for information in bond prices, we find that macroeconomic variables help predict short-maturity bond returns with statistical and economic significance, while the evidence for long-maturity bonds is much weaker. To understand this pattern, we provide a new decomposition of bond excess returns in terms of innovations of short-, medium- and long-run factors of the yield curve. We show that, in fact, macro data only contains unspanned predictive information about the short-run factor. This extra predictability varies over time and is stronger when inflation is high.

Keywords: Bond risk premia, yield curve, spanning hypothesis, Nelson-Siegel factors, machine learning

JEL Classification: G11, G12, G17

*We would like to thank Caio Almeida, Torben Andersen, Michael Bauer, Marcelo Fernandes, Andrei Gonçalves, Felipe Iachan, Leonardo Iania, Robert Korajczyk, Onno Kleen, Marcelo Medeiros, Geert Mesters, Humberto Moreira, Costis Skiadas, Viktor Todorov, Michel van der Wel, and Cynthia Wu for useful comments and suggestions. We also thank 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, and the Northwestern Kellogg Brown Bag Series for helpful discussions.
†freire@ese.eur.nl.
‡raul.riva@kellogg.northwestern.edu.

1 Introduction

Understanding the drivers of interest rates is fundamental for investors, researchers and especially policy makers. The simplest model of the term structure states that long-term yields are the average of future expected short-term yields. This is known as the *Expectations Hypothesis*, which equivalently posits that investors are risk-neutral and excess bond returns are not predictable. This hypothesis has been empirically rejected using information from the yield curve to predict bond returns, implying that bond risk premia are time-varying and important for explaining interest rates (Fama and Bliss, 1987; Campbell and Shiller, 1991; Cochrane and Piazzesi, 2005).

Whether the yield curve alone contains all the relevant information for predicting bond risk premia is a matter of more debate. Under certain assumptions shared by most macro-finance term structure models, the yield curve fully spans the information set of investors at time t. The *Spanning Hypothesis* therefore states that, controlling for the current yield curve, no other variables should help predict future yields and holding bond returns.[1] On the one hand, a number of papers use predictive regressions to show that macroeconomic variables have incremental predictive power for bond excess returns (Ludvigson and Ng, 2009; Joslin et al., 2014; Cieslak and Povala, 2015). On the other hand, Bauer and Hamilton (2018) cast doubt on the validity of the inference from the in-sample regressions used in these papers, while Bauer and Rudebusch (2017) demonstrate that such regression evidence is not necessarily inconsistent with the spanning hypothesis.

In this paper, we provide more nuanced evidence showing that the spanning hypothesis is in fact violated, but asymmetrically across bond maturities. Instead of considering in-sample regressions subject to the criticisms above, we focus directly on out-of-sample risk premia forecasts and assess whether incorporating macroeconomic variables on top of yield curve information results in better predictive ability. We show that macro data affords strong predictive power for future excess returns of bonds with shorter maturities (2 to 5 years), while for longer maturities the evidence is much weaker and the yield curve benchmark is rarely outperformed. In fact, the predictive power of macroeconomic data relative to the yield curve decreases monotonically with the bond maturity.

To help explain this pattern of violations, we provide a new decomposition of bond returns. Under the assumption that yields follow a dynamic Nelson-Siegel model (Nelson and Siegel, 1987; Diebold and Li, 2006), we show that bond excess returns can be written as a weighted sum of innovations of three factors capturing the short-, medium- and long-run behavior of the yield curve. The weights, which depend solely on the bond maturities, imply that the higher (lower) the maturity, the higher the relative importance of the long-run (short-run) factor in the bond return. This provides a natural framework for understanding asymmetric violations of the spanning hypothesis. Namely, we can investigate for *which* of these factors macroeconomic variables contain information not spanned by the yield curve.

We estimate the Nelson-Siegel factors for the whole sample and examine whether macroeconomic variables contain predictive power beyond the yield curve for each factor. We start by considering different sets of principal components summarizing the information of the macroeconomic data. We find that the principal components only help predict the short-run factor, while we cannot reject that the yield curve spans the relevant information for the medium- and long-run factor. As an alternative way of handling the high-dimensionality of the macroeconomic data, we also consider linear models with regularization. We focus on the Ridge (Hoerl and Kennard, 1970), Lasso (Tibshirani, 1996) and Elastic Net (Zou and Hastie, 2005). Across these different methods, macroeconomic variables again only matter for predicting the short-run factor. Importantly, our results are the same if we consider non-linear methods instead, such as the Random Forest (Breiman, 2001). These results shed light on the asymmet-

[1] The spanning hypothesis arises naturally in equilibrium models under rational expectations. See, for example, Wachter (2006), Piazzesi and Schneider (2007) and Rudebusch and Wu (2008). The typical argument is that, in equilibrium, prices should incorporate all possible information and be consistent with the true data generating process from the model. Since yields are an invertible function of traded bond prices, the argument follows. See Duffee (2013) for a review that discusses some of these models.

ric violations of the spanning hypothesis: macro data only contains unspanned predictive information about the short-run factor that affects disproportionately more short-maturity bonds than long-maturity bonds.

The asymmetric violations of the spanning hypothesis are not only statistically, but also economically significant. For a mean-variance investor optimizing her portfolio based on bond risk premia forecasts, exploiting information from macroeconomic variables generates Sharpe ratios around 50% higher relative to the yield curve benchmark for trading 2-year bonds. These improvements are statistically significant across different predictive models. In contrast, improvements are usually half as large for trading 10-year bonds and often not statistically different than zero. Again, the Sharpe ratio improvements decrease monotonically with the bond maturity. Since macroeconomic variables are only useful for predicting the short-run factor that more heavily influences bonds of shorter maturities, this result is the economic counterpart of the asymmetric statistical violations.

The findings described so far are unconditional in that they test whether there are violations of the spanning hypothesis on average. In order to understand not if but *when* these violations happen, we leverage the conditional predictive ability framework from Giacomini and White (2006). We find that the predictability of the short-run factor varies over time and depends on the inflation level. More specifically, periods of higher inflation are associated with periods in which the benefits from using macro data on top of the information from the yield curve are larger. One potential interpretation of this relation is that macroeconomic variables are informative about decisions of the Federal Reserve, whose instruments generally affect more the short end of the yield curve, and those decisions mainly occur as a response of periods of high inflation.

The remaining of the paper is organized as follows. After a brief review of the related literature, Section 2 describes the data we use. Section 3 studies the forecast of bond risk premia over different maturities. 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 analyzes the economic significance of violations of the spanning hypothesis, while Section 7 discusses the time-varying patterns of those violations. Finally, Section 8 concludes the paper.

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) show that different variables constructed using the yield curve are able to predict bond excess returns, which refutes the expectations hypothesis. Providing evidence against the spanning hypothesis instead, a number of 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). 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 the evidence of violations from in-sample predictive regressions. Cieslak (2018) argues that errors in investors' expectations about the short-rate induce ex-post bond return predictability. More recently, Bauer and Rudebusch (2020) 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 bond return predictability, while Hoogteijling et al. (2021) put forward factors based on yield changes that outperform factors based on yield levels for predicting bond excess returns. Borup et al. (2023) document that violations of the expectations hypothesis are state dependent and can be predicted by economic activity.

Our main contribution is to provide new out-of-sample evidence that violations of the spanning hypothesis are asymmetric in the dimension of the bond maturity. In particular, macroeconomic variables are more effective in predicting returns for bonds with shorter maturities than for those with longer maturities. We stress that our goal is not to draw a comprehensive comparison across different forecasting methodologies like Gu et al. (2020), Bianchi et al. (2021) and Medeiros et al. (2021). Instead, we show that our main empirical finding holds regardless of the forecasting method we use. Additionally, we emphasize that we compare forecasts done with and without macroeconomic variables, but always including information from the yield curve. In that sense, our benchmark is different from studies such as Bianchi et al. (2021) and Borup et al. (2023) and we are silent about the expectations hypothesis. Rather, we are concerned with testing whether the spanning hypothesis holds across bond maturities.

Second, we connect 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 in order to forecast the yield curve. With a similar goal, 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. Our goal is to help explain the asymmetric violations of the spanning hypothesis based on a new decomposition of bond excess returns implied by the dynamic Nelson-Siegel model.

Third, our paper is related to studies modeling the relation between macroeconomic indicators and the yield curve. Ang and Piazzesi (2003) model the joint dynamics of bond yields and macroeconomic variables in a Vector Autoregression imposing no-arbitrage. Diebold et al. (2006) use a Kalman-Filter coupled with the Nelson-Siegel representation for yields to jointly model these factors and macroeconomic variables in a more flexible way than Ang and Piazzesi (2003) but that is not necessarily arbitrage-free. Dewachter and Lyrio (2006) estimate a reduced-form model for the joint dynamics of inflation expectations, a few macroeconomic indicators and the term structure. One important difference of these studies from ours is that they focus on *jointly* modeling the yield curve and macroeconomic data, such that they are interested in matching the contemporaneous dynamics observed on the data and not necessarily a predictive relation.

Finally, we relate to papers investigating the economic content of bond return predictability. Thornton and Valente (2012) provide evidence that the economic significance of violations of the expectations hypothesis is weak, while 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 with a model incorporating stochastic volatility and unspanned macro 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. Borup et al. (2023) leverage the conditional predictive ability framework from Giacomini and White (2006) to document that violations of the expectations hypothesis are state dependent and can be predicted by economic activity. We use a similar conditional framework to show, instead, that violations of the spanning hypothesis for the short end of the yield curve are state dependent and can be predicted by inflation.

2 Data

We rely on two data sets for our empirical exercises. Our yield curve data comes from Liu and Wu (2021) while 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).

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). With respect to 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.[2] We start the sample in 1973 since the 10-year bond started being traded in late 1972.

In terms of 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 that were classified into eight different 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).[3] Table D.1 in Appendix D reports the full 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 B.1, in Appendix B, 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.[4] We follow the standard practice of this literature and use fully revised data.[5]

[2]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. The liquidity of these longer-term bonds over time has been a disputed topic, however. Hence, we adopt a similar strategy as Bianchi et al. (2021) in their investigation and focus on maturities up to 120 months.

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

[4]The St. Louis Fed also maintains a quarterly version of this data set, with even more macroeconomic series, called FRED-QD.

[5]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

3 Forecasting Bond Risk Premia

We start by investigating the predictability of holding bond returns in excess to 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 the standard framework to analyze bond return predictability, as in, for instance, Cochrane and Piazzesi (2005), Ludvigson and Ng (2009), Joslin et al. (2014), and Bianchi et al. (2021).[6]

More specifically, we let $y_t^{(n)}$ be the n-year zero-coupon yield at month t. Then, $y_t^{(1)}$ represents the 1-year risk-free rate at time t.[7] We denote by $xr_{t+12}(n)$ the excess return over the 1-year risk-free 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 forecasting $xr_{t+12}(n)$ should also help forecasting $y_{t+12}^{(n-1)}$.

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 very large, it is not feasible to simply 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. Under the spanning hypothesis, they should all be irrelevant for forecasting the excess bond returns, however. In the same spirit as Ludvigson and Ng (2009), we use principal components of the FRED-MD data set to summarize macroeconomic information.[8]

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 that summarizes 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.[9] Another strategy, building on Litterman and Scheinkman (1991), would be considering the first 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 instance, the forecast for January,

in empirical results.

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

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

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

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

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 generate a forecast for the returns that will be realized on January, 1990. As we move ahead in time, the amount of data used both in the extraction of principal components of the FRED-MD data set and the estimation of 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 A.1, 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 to forecast 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 goes slightly above in some cases. 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.

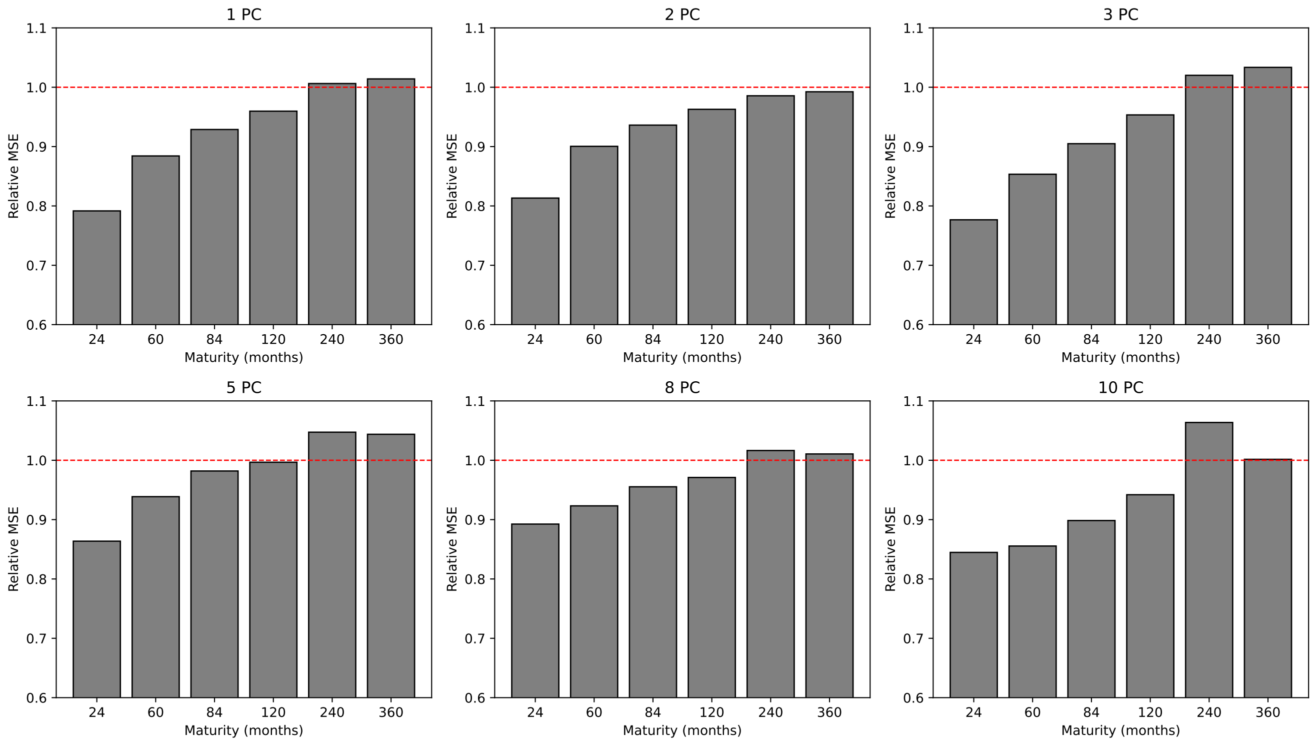
For this exercise, our preferred method for spanning the yield curve is using the forward rates, and that is exactly what we did in Figure 1. The reason is that forward rates are directly available from the zero-coupon yields and require no estimation. Nonetheless, as a robustness check, we repeat the out-of-sample forecasting exercise letting C_t stack the first three principal components of the yield curve, extracted in a sequential fashion as done with the macroeconomic data. Results are reported in Figure B.2 in Appendix B. The same asymmetry arises, where the absolute decrease in the MSE is even larger for shorter maturities.[10]

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.[11] We let C_t stack the forward rates. The p-values for the test are reported in Table 1. The null hypothesis is equal predictive ability, which would be implied by the spanning hypothesis.

[10] 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 has not 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.

[11] We allow for autocorrelation in the forecasting errors and deploy the HAC estimator from Newey and West (1987) to compute the variance required by Diebold and Mariano (1995).

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. Principal components are extracted in real time and do not introduce any look-ahead bias.



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.

4 Decomposing Bond Risk Premia

Taken at face value, the empirical results in the previous section suggest that macroeconomic variables can enhance the forecast of excess bonds returns, but in an asymmetric way: they are more helpful for the shorter end of the yield curve than for the longer end. We find evidence of this asymmetry both in-sample and out-of-sample. The latter approach is the focus of this paper.

We now develop a methodology that will enable us to interpret and quantify this finding more thoroughly. The

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

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.[12] 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 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 in 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.

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

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 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 new decomposition of the excess bond returns. From now on, we assume a constant $\lambda_t = \lambda > 0$ since that will be part of our identification 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:*

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

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 the 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, a simple 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).[13]

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

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

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 30-month horizon. Additionally, it facilitates comparisons with other studies that followed the same methodology.

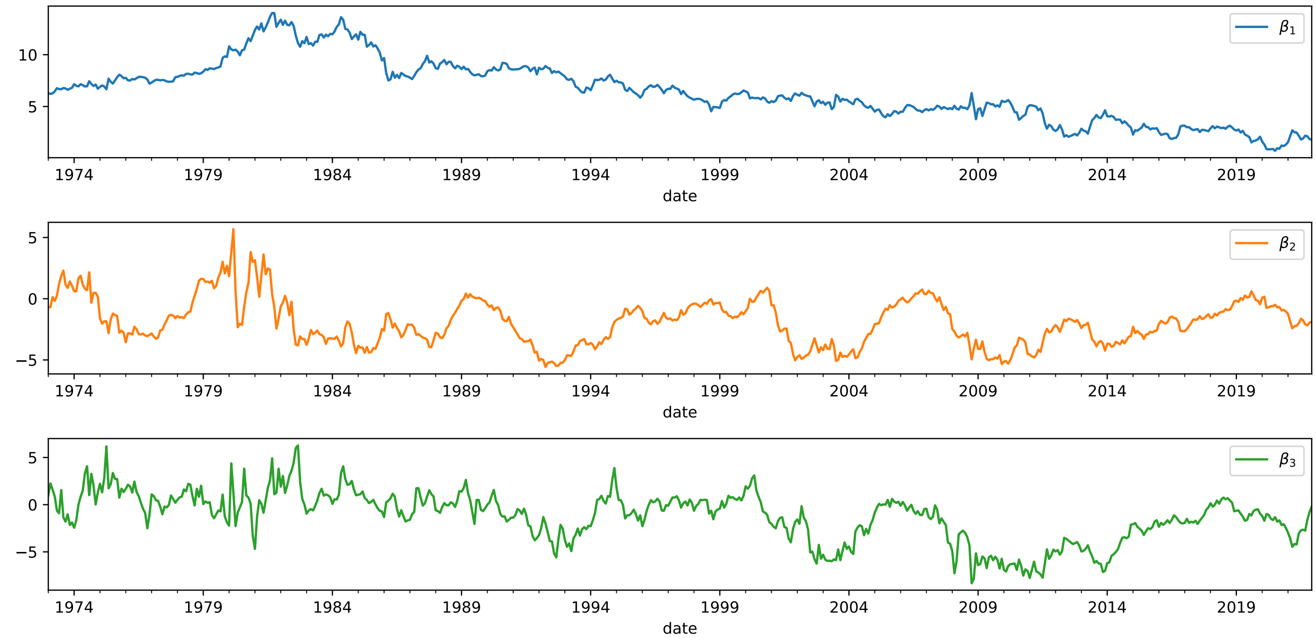
Table 2: Summary statistics for the estimated Nelson-Siegel factors. For each date, factors are estimated running a linear regression of observed yields on the loadings. We use all yields from 12 to 120 months that are available from the data set provided by Liu and Wu (2021). The sample ranges from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

[13] 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 in 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.

Figure 2: Estimated factors using the OLS approach of Diebold and Li (2006), with $\lambda = 0.0609$. For each date, factors are estimated 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.



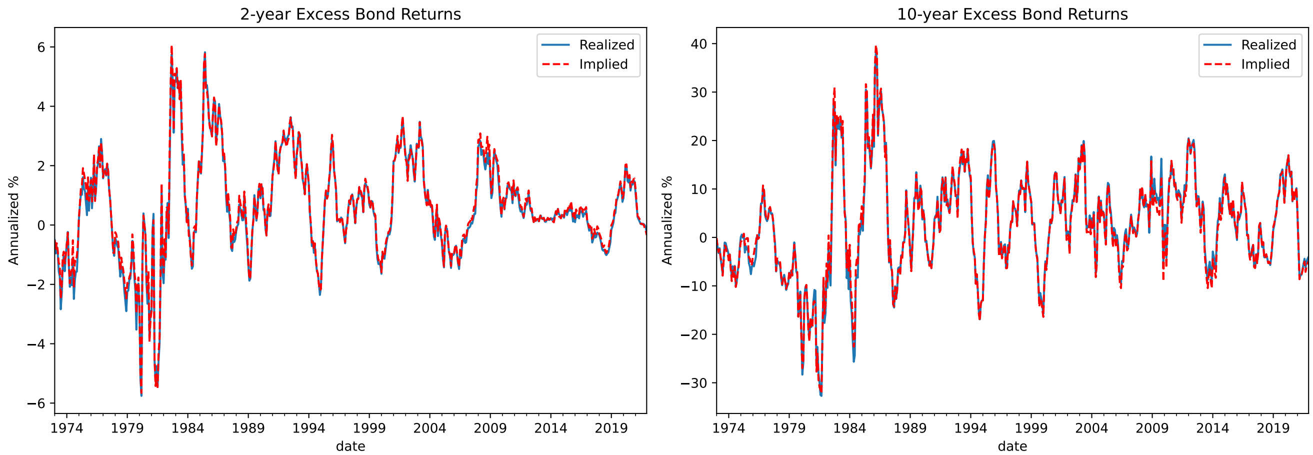
also on a linear time-trend. At the 5% level, we reject the hypothesis of an unit root both for β_2 and β_3, for both versions of the test. For the long-run factor, however, we only reject the null at the 10% level when we also control for a linear time trend. In general, the long-run factor is much 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 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 3 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 a 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.

Figure 3: 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 2.



In any case, 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.

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, all information needed to forecast these factors should be spanned by the zero-coupon yields themselves. One way to test the spanning hypothesis is asking whether macroeconomic data can enhance the forecast of these factors. A positive answer represents evidence against such hypothesis. Additionally, 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. As we did with bond risk premia, we focus on out-of-sample predictive exercises.

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 an expanding window forecasting exercise, as we did in Section 3. Our first forecast is made for January 1990 and the last one for December 2021, for a total of 384 out-of-sample forecasts. For example, for the January 1990 forecast, we only use data available up to January 1989. This data is used both for the extraction of the principal components from the FRED-MD data set and for the estimation of the linear model in (8). After the model is estimated, we compute the 12-month-ahead forecast for each of the factors.

We emphasize two aspects of this forecasting design. First, for each new forecast for the realizations taking place at $t + 12$, we perform the principal components extraction with data only up to time t. An econometrician in 1989, for instance, furnished only with a truncated version of our data, would have extracted the same principal components as we did. Similar expanding window designs have been recently used by Gu et al. (2020) and Bianchi et al. (2021). Second, the expanding window design is useful because the amount of data for estimation increases at each step, but the forecasts will have "long-memory": the dynamics of the beginning of the sample will affect forecasts at the end of the sample. We will later investigate a rolling window design and show that our conclusions do not depend on an expanding window.

We follow the tradition in the forecasting and Machine Learning literature and focus on the out-of-sample R^2 of our predictions:

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

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 fairly persistent time-series.[14] It 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:

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

We then infer the predicted level by:

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

Finally, we also stress that, under the random walk benchmark, Proposition (1) still implies a time-varying risk premium. This is a richer baseline than just assuming it is a constant value, possibly depending on the maturity - the so-called *expectation* hypothesis.[15] Our baseline case assumes that risk premium can vary over time since the factors themselves are allowed to change over time.

[14]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.

[15]See Bianchi et al. (2021) and Borup et al. (2023) for recent studies regarding the expectations hypothesis.

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 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.

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 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

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 numbers of principal components taken from our macroeconomic data, we 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 A, Table A.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 macro macro data. Alternatively, Tables A.3 and A.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 spanning hypothesis are stronger at the shorter end of the yield curve.

5.2 Regularized Linear Models

Our evidence so far uses principal components of the macroeconomic variables to summarize the information spanned therein. 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.[16] 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

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

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 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.[17] 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.

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.[18]

As we did in our out-of-sample exercise with principal components, we adopt an expanding window framework. Our first forecast is made for January 1990 and our last one is made for December 2021, such that we compute 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 of course associated to the specific pair (ψ_1, ψ_2) then used. We minimize the mean squared-error in the validation set picking the best possible candidate combination (ψ_1, ψ_2) from an user-specified grid using a simple grid-search method. We let the validation set represent 20% of the data available at a given

[17] 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.

[18] For a deeper discussion of this interpretation, see Hastie et al. (2009) and Giannone et al. (2021).

point 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.[19]

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 we B 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.

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 displays 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 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

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

[19] 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.

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 controlling 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.[20]

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 has to do with the fact that these factors are persistent and regularized models are known to work empirically worse in such scenarios.[21] 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. While the analysis using principal components was virtually silent regarding what variables were more important for forecasting, the regularized models are more explicit in that front. Since we have to numerically solve (12) 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.[22]

We compute how frequently variables from each group were chosen and show results in Figure 4 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 in the

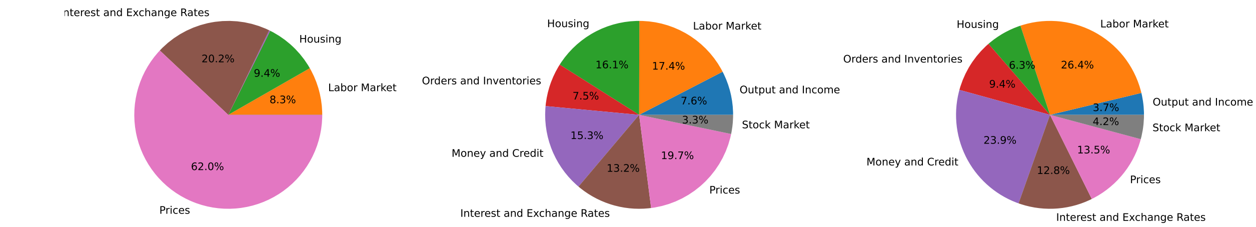
[20] Our (unreported) results show that, on average for both Lasso and Elastic Net, the typical chosen model here uses around 15-25 variables from the FRED-MD data set, completely ignoring the other ones.

[21] Figure B.3 in Appendix B further illustrates this point.

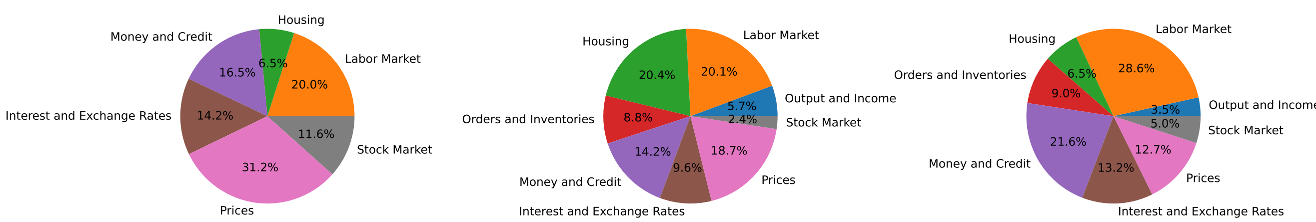
[22] See our Appendix D for a full list of variables and their classification.

Figure 4: 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 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 ($\beta_1, \beta_2, \beta_3$)



pie charts. The pie charts on the first row show results for the Lasso and the ones on the second row for the Elastic Net. For β_1, we show results for the innovations due to the poor performance when targeting its level, as shown in Figure B.3. 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.[23]

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 CPI measures, consumption expenditure indexes and commodity prices. We see these variables as all linked to the current state of inflation. This result seems intuitive for 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.

The dominance of these price indicators is less extreme when we use the Elastic Net in contrast to the Lasso, although these indicators are the prevalent ones in both cases. This is also expected. Heuristically, the L^2-norm

[23] 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.

penalty from the Elastic Net acts shrinking all coefficients towards zero while the L^1-norm penalty forces some of them exactly to zero. Since the coefficients have already been shrunk, the cost of allowing them to be non-zero is relatively smaller. Therefore, Elastic Net will tend to pick more variables than the Lasso.

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 don't take this result as a surprise, though. 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 which are informative about the current state of the economy helps forecasting this factor. The more puzzling fact is why this information is not already contained in the current yield curve itself, spanned either by the lagged factors or the forward rates, i.e., why the spanning hypothesis fails for the shorter end of the curve. Giannone et al. (2021) analyze how these regularization methods perform when making forecasts in different contexts in Economics and find 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. A similar pattern arises for β_3, for which our methodology ends up mixing the different signals available in the data set.

5.3 Non-Linear Methods

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 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 allow for non-linearity in two ways. The first one uses higher powers of principal components while the second one employs a Random Forest method.

5.3.1 Powers of Principal Components

Our first attempt to study whether non-linearities are important for forecasting in this environment revisits the forecasting equation (8) allowing for higher-order powers of principal components. Given a number K of principal components to use, we 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:

$$PC_t^{(r)} \equiv \begin{bmatrix} PC_{1,t} & \cdots & PC_{K,t} & PC_{1,t}^2 & \cdots & PC_{K,t}^2 & \cdots & PC_{1,t}^r & \cdots & PC_{K,t}^r \end{bmatrix}'.$$

We use this richer version of the principal components in the following forecasting regression:

$$\beta_{i,t+12} = \alpha_i + \theta_i' C_t + \gamma_i' PC_t^{(r)} + \epsilon_{i,t+12}, \quad i \in \{1, 2, 3\}. \tag{14}$$

We report results for $r = 2$ on Table 5 and leave results for $r = 3$ for Table A.6 in Appendix A. We proceed as before, using the period 1990-2021 as our out-of-sample sample, and extracting principal components as we increase the sample following the expanding window strategy. For the sake of space, we focus on results using $K \in \{3, 5, 8\}$.

Panel A in Table 5 displays results controlling for the information contained in the forward rates. The same story from previous exercises remains, where we can only both improve on "No Macro" baseline and beat the random

Table 5: We report the out-of-sample R^2 from the forecasting model in (14). The out-of-sample period is 1990-2021. We set $r = 2$ 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, $r = 2$							
	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

walk for β_2. This is true when we directly target the level of the factors and when we target the innovations. Interestingly, the absolute forecasting accuracy is similar to the previous $r = 1$ case.

Panel B repeats this exercise controlling for the lagged Nelson-Siegel factors. Once more, we can only outperform the benchmark for β_2. In this case, the "No Macro" baseline already beats a random walk, but the model using the principal components from the macroeconomic data set increases the out-of-sample R^2 threefold, which once more represents a strong violation of the spanning hypothesis.

The results for $r = 3$ in the Appendix also display the same pattern. However, with $r = 3$ and $K = 5$, for example, we have 15 extra parameters to estimate with the same amount of data, aside from the 10 parameters from the baseline model when we control for forward rates. This is translated to lower statistical power, as can be seen from the larger p-values.

5.3.2 Random Forest

Although equation (14) allows for non-linear transformations of principal components, the forecasting equation is still linear. We take it further and implement a fully non-linear method: the Random Forest. 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 similar fashion, Goulet-Coulombe (2023) does an in-depth analysis of their use (and success) when predicting macroeconomic variables.[24]

[24] 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.

Table 6: 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

To implement this model, we adopt the standard CART algorithm from Breiman (2001).[25] 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.[26] At each time t, we reestimate the forest and predict both factors and their innovations one year ahead.

Table 6 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. 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.

The most important message from Table 6 is, however, that the asymmetric violations of the spanning hypothesis we previously documented are not due to our previous choice of linear forecasting methods. We only find violations through β_2 no matter how we approach the problem.

5.3.3 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

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

[26]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).

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 for the in-sample MSE reduction, i.e., how helpful each variable was in predicting the Nelson-Siegel factors.

We classify the variables in two classes: the ones that come directly from the yield curve data and the ones that come the FRED-MD data set. Figure 5 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). This behavior is consistent with the spanning hypothesis, under which no information provided by the extra state variables should help with forecasting. On the other hand, the behavior for β_2 is exactly the opposite. More than 90% of the MSE reduction is attributed to the macro 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 6, 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 hypothesis.

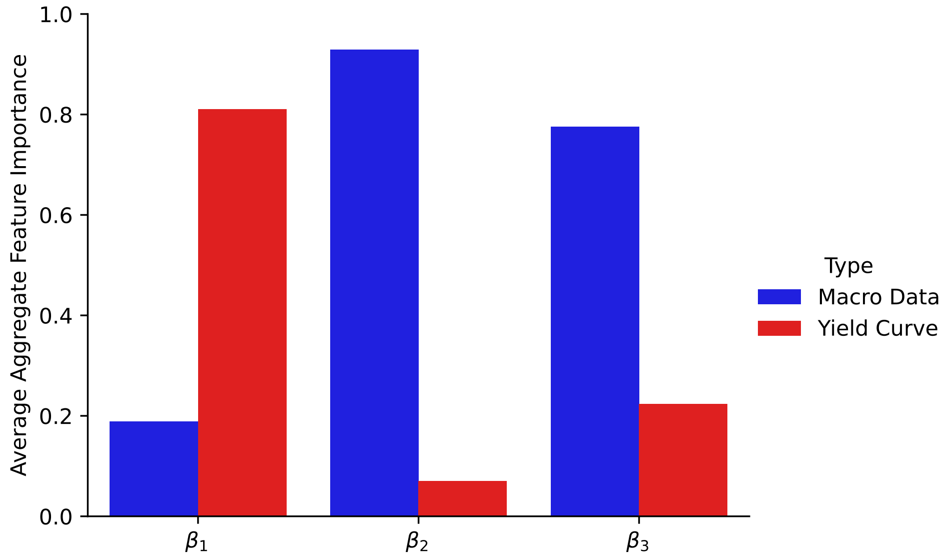
Figure B.4 in Appendix B displays the average feature importance F_i for each variable and each factor, keeping the same color coding. In line with the aggregation reported in Figure 5, the red bars dominate the plot for β_1, while they are very small for β_2. Additionally, one can see that there is 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.

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 Economic Significance of Violations

So far, we have established that violations of the spanning hypothesis are stronger and, in fact, only seem to occur at the shorter end of the yield curve. Now, we assess whether these violations are economically meaningful. We present a setup in which a consumer that 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).

Figure 5: 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 importances from the FRED-MD dataset. The out-of-sample period is 1990-2021.



6.1 Setup

We study the problem of a consumer that 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 $R_{p,t}(\boldsymbol{w}_{t-12}) = 1 + y^{(1)}_{t-12} + \boldsymbol{w}'_{t-12}\boldsymbol{xr}_t$, where $\boldsymbol{w}_{t-12}$ is a vector of portfolio 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 risk premia, respectively. Given these two objects, we know the solution for the problem above in closed form:

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

Our methodology readily delivers estimates of $\boldsymbol{\mu}_{t+12|t}$ with and without macroeconomic data through different methods. We do not have, however, a full model for the covariance matrix, which is not the focus of this paper. 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 for 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. For the forecasts incorporating information from macroeconomic variables, we focus on our baseline linear specification with principal components.

6.2 Sharpe Ratio Improvements

We keep track of the investors 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 $A = 3$.

The Nelson-Siegel factors were useful, when coupled with Proposition 1, to perform a formal assessment of the asymmetry displayed in Figure 1. But when trading, we are obviously concerned with the total risk premia earned by the consumer. Hence, we revisit the empirical exercise summarized by equation (2) and use both the time-series of forecasts with and without macro data to construct the conditional risk premia and the conditional covariance matrix of risk premia.

Our simple setup abstracts away from trading costs related to leveraged positions or even short positions in some of these bonds. In order to bring the setup closer to what a practitioner would do in practice, 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:

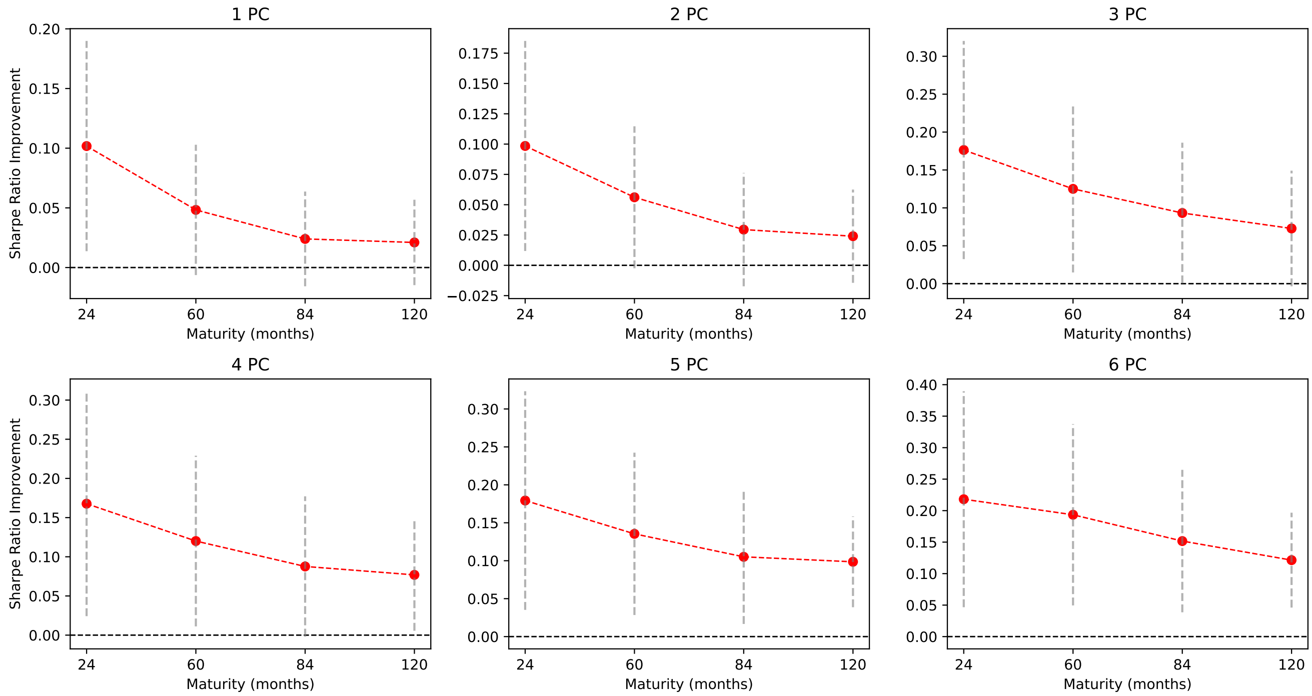
$$Improvement \equiv SR^{(\text{Macro})} - SR^{(\text{SH})}.$$

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. Figure 6, which uses the forward rates as controls for the yield curve, displays the improvement across different maturities (2, 5, 7, and 10 years) of using different numbers of principal components of macroeconomic variables for the more strict case constraining portfolio weights.

The solid dots represent the difference between achieved out-of-sample Sharpe ratios. The grey dotted bars are 95% confidence bands computed using the asymptotic approximation from Ledoit and Wolf (2008). The profile of these Sharpe Ratio improvements is decreasing as a function of maturity, regardless of the included number of principal components. In fact, when we inspect the improvement for the 10-year bond, for example, it is frequently not statistically different than zero. On the other hand, the confidence bands never include zero for the 2-year bond.

Taking the point estimates at face value, the improvements for the shortest maturity considered range from 0.10 to 0.21. We consider these improvements as far from being trivial. The decreasing shapes from Figure 6 should not be

Figure 6: 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.



taken as a surprise in light of our previous results and in light of Proposition (1). Since macroeconomic variables are useful only for forecasting the short-run factor β_2 and the shorter maturities are more heavily influenced by movements in β_2, this result is the manifestation of the asymmetric violations in the context of trading.

Figure B.5 in Appendix B repeats the trading exercise, now without shorting and leverage constraints. The same decreasing shapes and overall qualitative results stand, showing that these asymmetric improvements in Sharpe ratios are not due to our strict constraints. The absolute value of the improvements is also larger since the consumer has greater freedom when trading. Nonetheless, one should take these point estimates more as upper bounds of the increase in Sharpe ratios one can achieve, taking into account that we do not model trading costs, margin calls, and other practical details.

7 When Are Violations Happening?

After providing robust evidence that violations of the spanning hypothesis occur asymmetrically across the yield curve and that this phenomenon is economically meaningful, we explore these violations over time. Concretely, 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 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:

$$D_{i,t} \equiv L_{i,t}^{(\text{SH})} - L_{i,t}^{(\text{Macro})}, \tag{15}$$

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

$$H_0 : \mathbb{E}\left[D_{i,t+12}|\mathcal{G}_t\right] = 0. \tag{16}$$

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 (16). 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:

$$W \equiv T \cdot \mathbf{z}_{t+h}'\widehat{\Omega}_T^{-1}\mathbf{z}_{t+h} \xrightarrow{d} \chi_q^2. \tag{17}$$

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.

7.1 Evidence from Rolling Windows

One drawback from this methodology, as emphasized by Giacomini and White (2006), is that it can only be applied to *rolling window* estimation schemes. Hence, we cannot directly apply their test to our current forecasting setup since we were using expanding windows. Instead, we design a fixed, rolling window forecasting exercise. This serves two purposes. First, it ensures that our previous findings are not due to the choice of the estimation window. Second, it generates time-series for $D_{i,t}, i = 1, 2, 3$ that will be further used in a conditional predictive ability test.

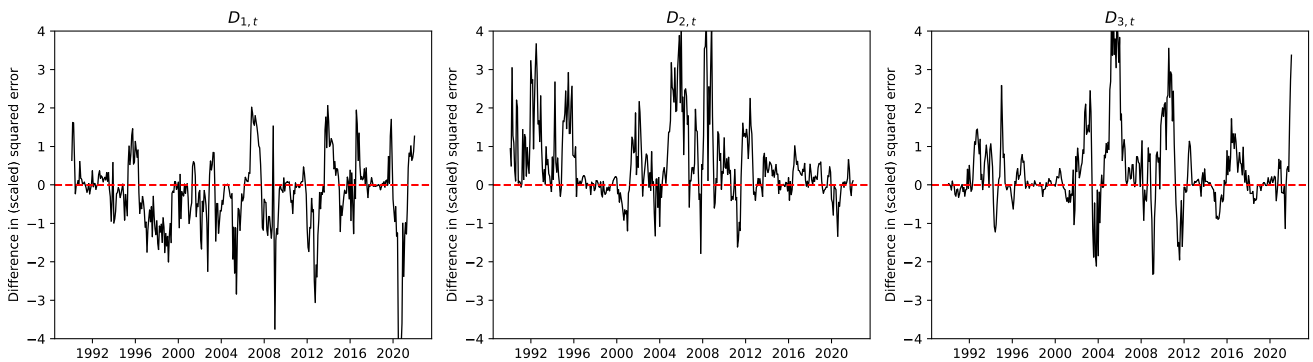
The fact that we now use fixed windows poses numerical challenges for all our methods. Regressions using principal components will suffer both because there is less data to estimate a linear regression but also because there is less data to extract principal components from the panel of macroeconomic variables. Similarly, the cross-validation technique used on the estimation of the regularized models would be rather imprecise with just a few data points to pick suitable values for (ψ_1, ψ_2). Then, we focus on results using the Random Forest method. Aside from being the one with the best forecasting power in Section 5, it is known to behave well with limited amounts of data.[27]

There is also 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, but also show results with a 120-month window as robustness.

We keep the same out-of-sample period from before (1990-2021) and forecast the level of the factors as we did before, with the important difference that each forecast is done with a fixed number of data points, both under the spanning hypothesis and when conditioning on the full set of extra state variables.

Figure 7 reports the realizations of $D_{i,t}$ when controlling for the forward rates. 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.

Figure 7: 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.



We see that D_1 lives mostly on the negative plane, while D_2 is situated mostly in the positive plane. This is consistent with our previous evidence that macroeconomic data was helpful for forecasting β_2, but not β_1. Once more, the behavior for β_3 is more mixed. Interestingly, we also see that the extra predictive power brought by the macroeconomic variables is concentrated in a small time frame. For example, macro data seemed especially helpful during the 1990s and early 2000s for β_2. Figure B.6 in Appendix B displays the analogous time-series controlling for the lagged Nelson-Siegel factors, displaying the same qualitative behavior.

We first test whether we can reject that $\mathbb{E}[D_{i,t}] = 0$, i.e., we use the trivial sigma-algebra. Table 7 reports the out-of-sample R^2 for a 180-month window, controlling for the lagged Nelson-Siegel factors on the left and by the forward rates on the right. The inclusion of macroeconomic data did not improve the forecast for β_1 in neither case. However, there is a large improvement for β_2, which makes us beat both the "No Macro" baseline and the random walk. There is also an improvement for β_3, but the model is never better than a random walk.

The results in Table 7 reassure us that our previous evidence is not due to the expanding window nature of our forecasts. Table A.7 in Appendix A reports results for this exercise with a 120-month window and tells the same

[27]For a discussion of the flexibility associated to regression trees and Random Forests, see Hastie et al. (2009) and Goulet-Coulombe (2023).

Table 7: 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.

	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

story. The overall predictability is a bit lower, which is expected since the Random Forest is estimated with less data at each point in time. In summary, the evidence from the rolling window forecast agrees with our previous findings. We can only spot violations of spanning hypothesis through the prediction of the short-run factor β_2.

7.2 Conditional Predictive Ability

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 dynamic term structure models provides no guidance here. 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 unemployment rate for the U.S. and its natural unemployment rate, both reported by the St. Louis Fed. The fact that the predictability we find stems from the short-term factor β_2 suggests that phenomena related to predictability might be associated to monetary policy decisions taken by the Federal Reserve. Its dual mandate takes unemployment into consideration so we add it here as well.

4. Year-over-Year inflation (PCE): it is the 12-month inflation measured by the Personal Consumption Expenditure excluding energy and food prices, also taken from the St Louis Fed. Our motivation here is the same as UGAP since inflation control is one of the targets of the Federal Reserve and the current state of inflation is relevant for monetary policy decisions. Additionally, as emphasized by Bauer and Rudebusch (2017), UGAP and PCE are usually included as risk factors in some traditional dynamic term structure models.

5. Slope: we define this as the difference between the 10-year rate and the Fedfunds 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 B.7 in Appendix B. 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 8 reports at its last row the p-values for the conditional predictive ability test based on (17). 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.

Table 8: 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 7. Each column also reports the estimate for $\boldsymbol{b}$ from (18). 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

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:

$$D_{2,t+12} = a + \mathbf{b}'\mathbf{x}_t + u_{t+12}, \tag{18}$$

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 to moments where the gain in using macroeconomic variables to forecast ahead was higher.

Column 6 uses the variables used in Borup et al. (2023). But with a p-value of 0.50 we cannot reject the null hypothesis at 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 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 8 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 $D_{2,t+12}$.

Although Table 8 spans the information from the yield curve using the forward rates, our result regarding the PCE does not depend on that. Table A.8 in Appendix A tells a similar story, although the estimated coefficient is around 0.25. 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. In other words, violations of the spanning hypothesis in the short end of the yield curve tend to occur when inflation is high.

7.3 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 12-month inflation within each inflation tercile was 1.3%, 1.8% and 2.8%, respectively.

For each tercile, we compute the average value for $D_{i,t}, i = 1, 2, 3$. The results are reported in Table 9. 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.

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 9 is the same as the one we took from the more formal analysis of Table 8: moments of higher inflation at time t are associated with more likely violations of the spanning hypothesis through β_2 in our sample.

Table 9: 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

8 Conclusion

We document violations of the spanning hypothesis that are asymmetric 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 conclusion. We also show that a mean-variance consumer would benefit from using macroeconomic variables when trading bonds in an asymmetric way. The benefits from extra predictability are larger, for example, when trading the 2-year rate than when trading the 10-year one.

Using a conditional predictive ability test, we provide evidence that periods of higher inflation are associated with periods in which the benefits from using macro data is larger, no matter how we control for the information already contained in bond prices. This correlation makes us conjecture whether the effect we document is related to monetary policy. Moments of high inflation are moments in which the Federal Reserve will be more prone to act, all else equal. Their monetary policy instruments affect, in general, the shorter end of the yield curve more than the long end. If macroeconomic data helps forecasting the decisions of the monetary authority, it should also help forecasting yield movements for shorter maturities. This effect is ruled out by standard term structure models that typically assume a *known* mapping between risk factors and the short rate, which is equivalent to assuming that the policymaker's reaction function is known.

In summary, we argue that the answer to the question of whether the spanning hypothesis is true or not is far from being simple and binary. Instead, we document a more nuanced landscape, with large asymmetries in its violations across maturities. From a theoretical perspective, an important challenge is finding micro-foundations for the asymmetries we document and their subsequent predictability. That is a non-trivial task since the typical equilibrium models we use in macro-finance imply the spanning hypothesis across the entire yield curve. We believe that models that step out of the FIRE (full-information rational expectations) framework represent a fruitful line of future research.

References

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.

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.

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

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.

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.

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

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.

Dewachter, H. and Lyrio, M. (2006). Macro factors and the term structure of interest rates. *Journal of Money, Credit and Banking*, 38(1):119–140.

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

Duffee, G. R. (2013). 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.

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.

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.

Hoogteijling, T., Martens, M. P., and van der Wel, M. (2021). Forecasting bond risk premia using stationary yield factors. *SSRN Electronic Journal*.

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.

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.

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

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.

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. and Schneider, M. (2007). Equilibrium Yield Curves. In *NBER Macroeconomics Annual 2006, Volume 21*, NBER Chapters, pages 389–472. National Bureau of Economic Research, Inc.

Rudebusch, G. D. and Wu, T. (2008). A macro-finance model of the term structure, monetary policy and the economy. *The Economic Journal*, 118(530):906–926.

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

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.

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.

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

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 Tables

A.1 In-sample Evidence - Bond Risk Premia

Table A.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 start 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 A.2 in the same appendix. 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.

Table A.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
R2 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
R2 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 A.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 A.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
R2 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
R2 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

A.2 In-sample Evidence - Nelson-Siegel Factors

Table A.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-squared 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-squared 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 A.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 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	-0.06**	-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-squared 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-squared 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

A.3 Additional Out-of-sample Evidence

Table A.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 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 A.6: We report the out-of-sample R^2 from the forecasting model in (14). 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
Beta 1	-0.21	-0.15	-0.13	-0.04	0.19	0.20	0.04
Beta 2	-0.08	0.13	0.09	0.03	0.09	0.15	0.30
Beta 3	-0.12	-0.10	-0.13	-0.19	0.35	0.56	0.81
Innovation 1	-0.19	-0.12	-0.11	-0.02	0.15	0.17	0.03
Innovation 2	-0.11	0.12	0.08	0.03	0.08	0.13	0.25
Innovation 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
Beta 1	-0.10	-0.12	-0.09	-0.02	0.67	0.47	0.14
Beta 2	0.06	0.13	0.11	0.05	0.33	0.37	0.53
Beta 3	-0.11	-0.09	-0.14	-0.20	0.40	0.70	0.83
Innovation 1	-0.10	-0.12	-0.09	-0.02	0.67	0.47	0.14
Innovation 2	0.06	0.13	0.11	0.05	0.33	0.37	0.53
Innovation 3	-0.11	-0.09	-0.14	-0.20	0.40	0.70	0.83

Table A.7: 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

Table A.8: 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 B.6. Each column also reports the estimate for $\boldsymbol{b}$ from (18). 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)
spread					0.26**				0.18	0.15
					(0.13)				(0.12)	(0.12)
N	384	384	384	384	384	384	384	384	384	384
R2	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

B Additional Figures

Figure B.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.

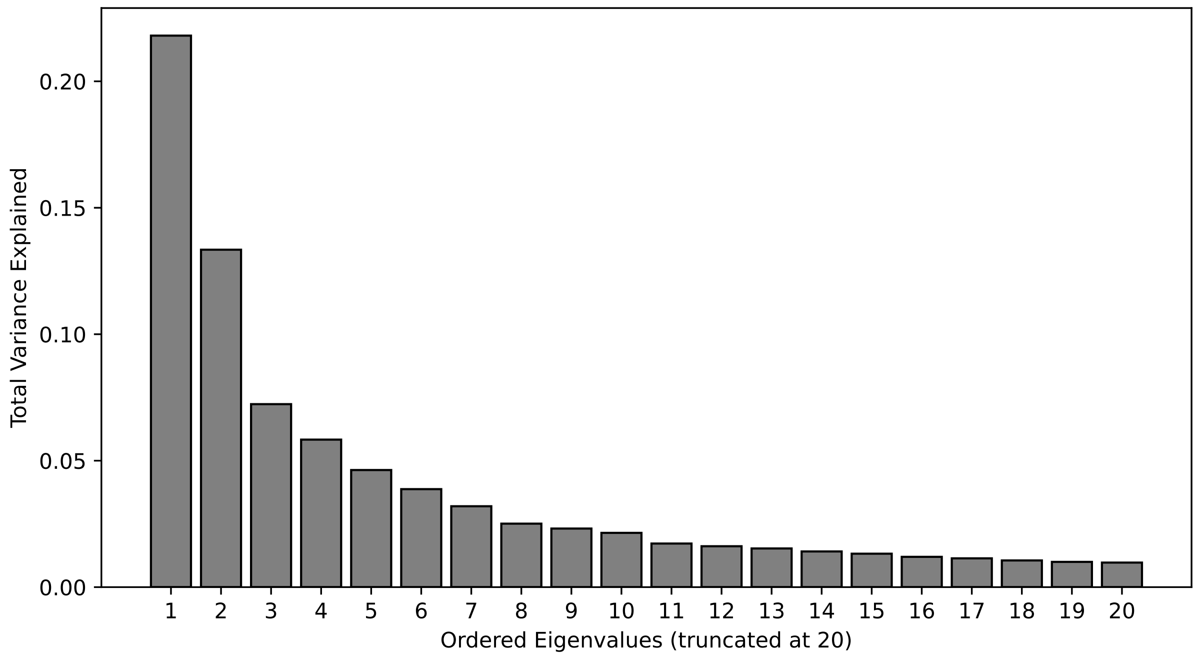


Figure B.2: Relative MSE predicting returns using three principal components of the yield curve as controls. For each maturity we show the ratio between the MSE attained with different numbers of principal components from the macroeconomic data and the baseline model that uses information only from the 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. Principal components are extracted in real time and do not introduce any look-ahead bias.

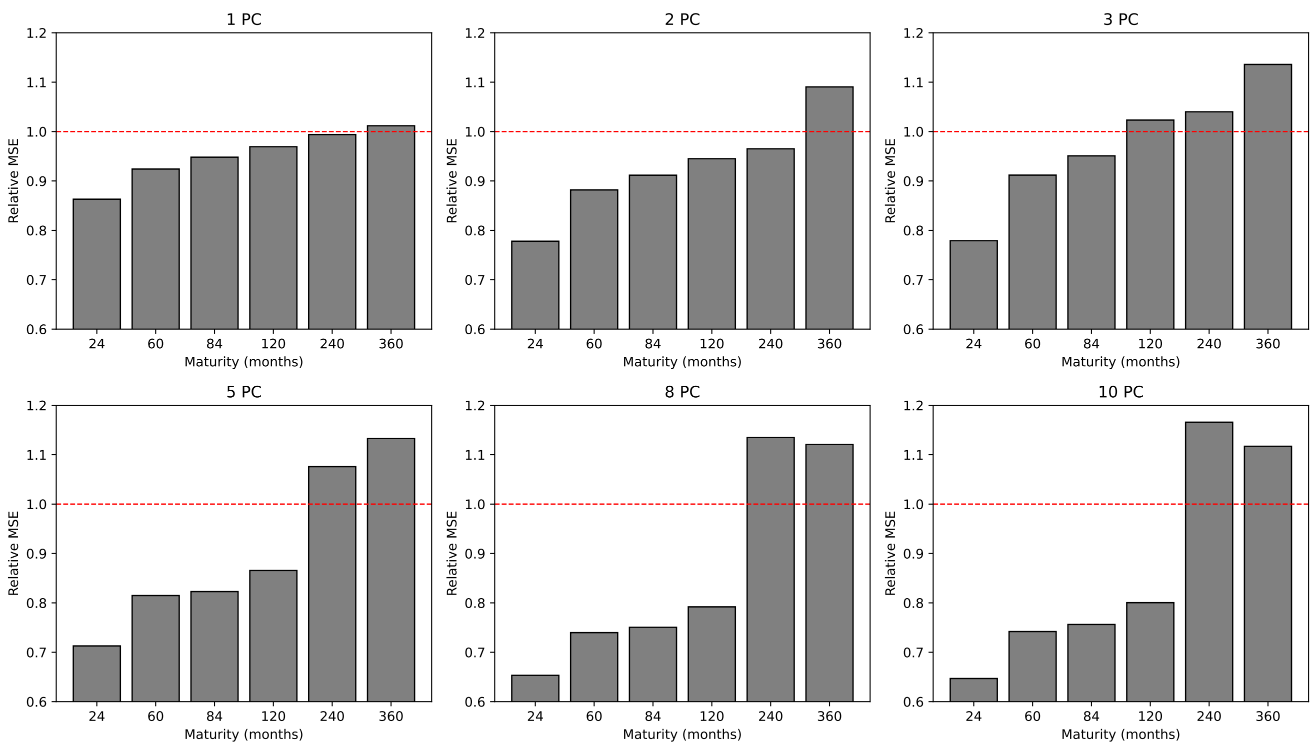


Figure B.3: Forecasts for β_1 both targeting its level (left) and its innovation (right) created by the ElasticNet. The out-of-sample period is January 1990 to December 2021. The black dashed line represents the realizations we estimated as shown in Figure 2. The forecasts for the level show a large 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 2). 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 the innovations, on the other hand, 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.

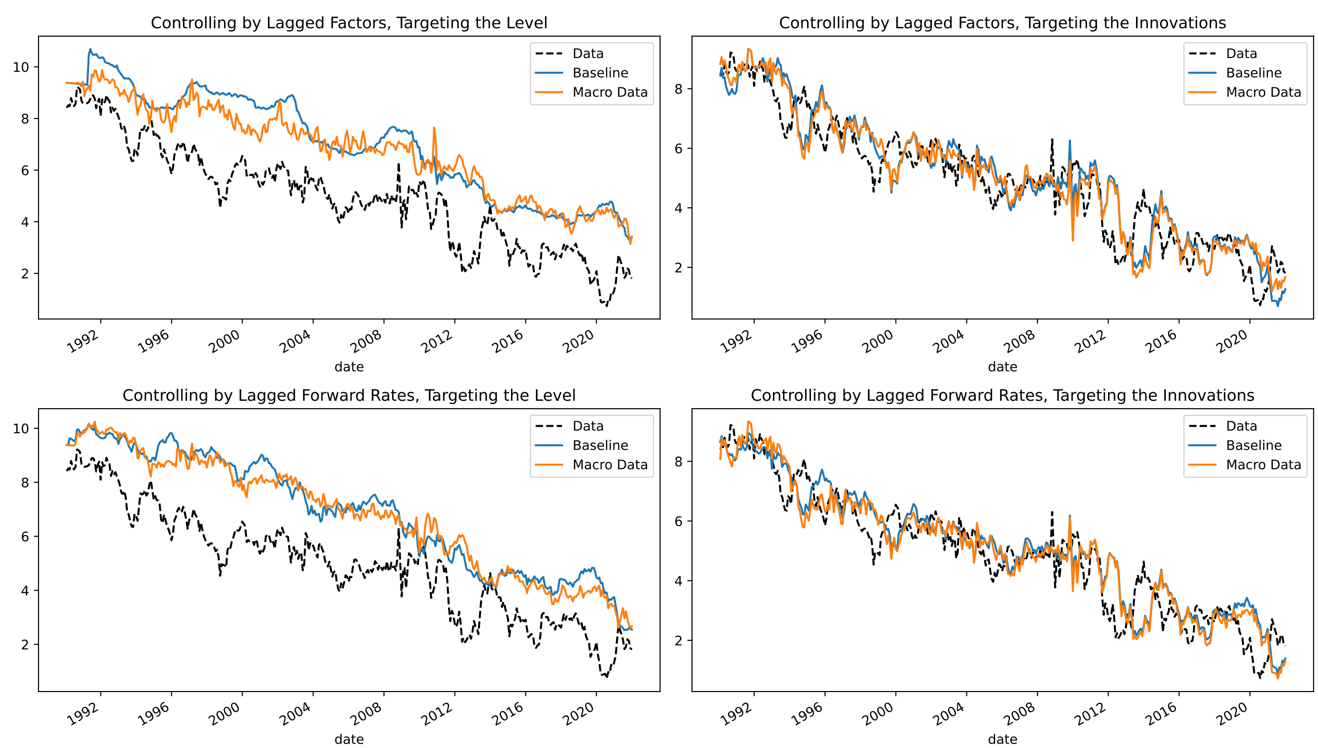


Figure B.4: Individual feature importance for each variable used in splits for the Random Forest strategy. We compute the feature importance of each variable for each tree and then average over trees and over time, normalizing everything at the end. Blue bars represent variables from the FRED-MD dataset 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.

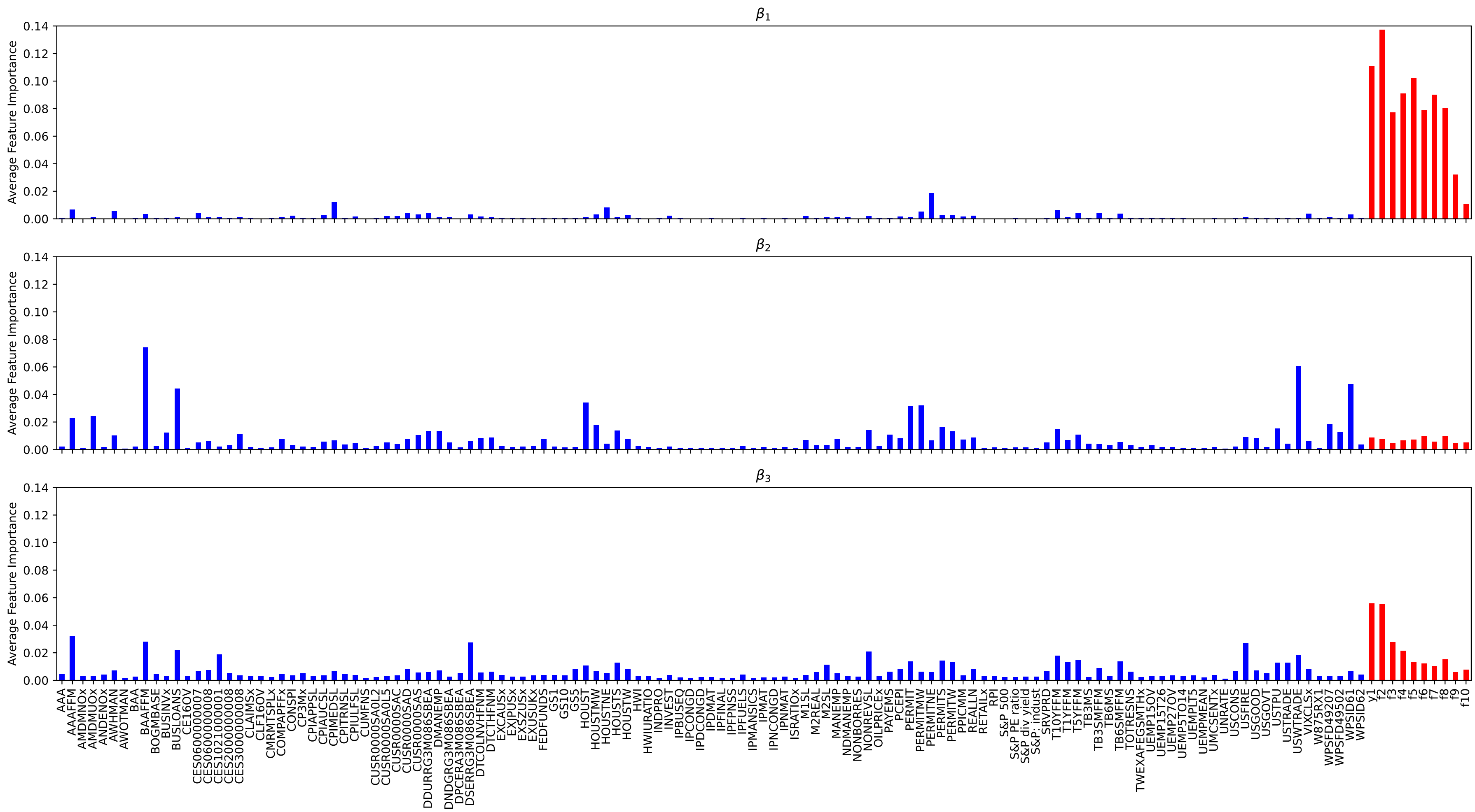


Figure B.5: This figure reports the Sharpe ratio improvement the mean-variance consumer can achieve when using principal components of the FRED-MD dataset to trade bonds of different maturities. The out-of-sample period is 1990-2021. The dots represent point estimates while the gray bars represent 95%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

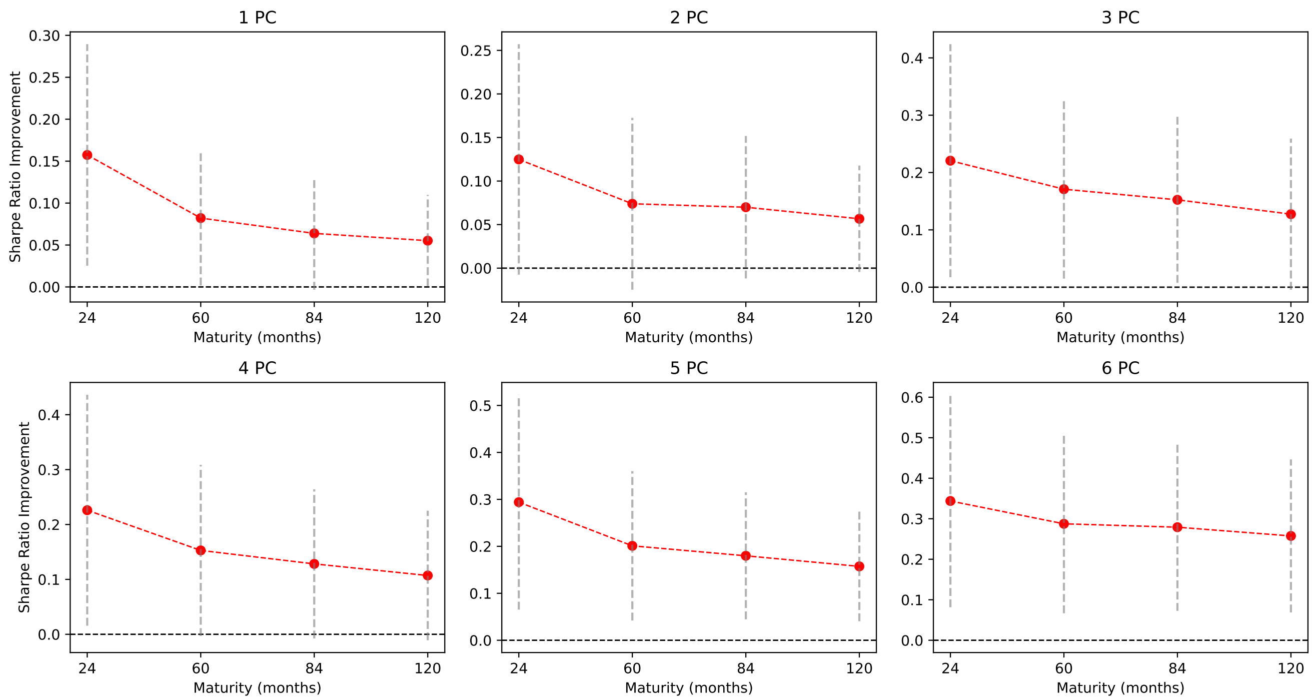


Figure B.6: This figure shows time-series for $D_{i,t}$ estimated using a Random Forest with 500 trees. The out-of-sample period is 1990-2021. Here we control for the lagged Nelson-Siegel factors and use a 180-month rolling window for estimation. All three series are scaled by their respective standard deviation so the scales are comparable. A positive value means that macro data was useful when forecasting the Nelson-Siegel factors.

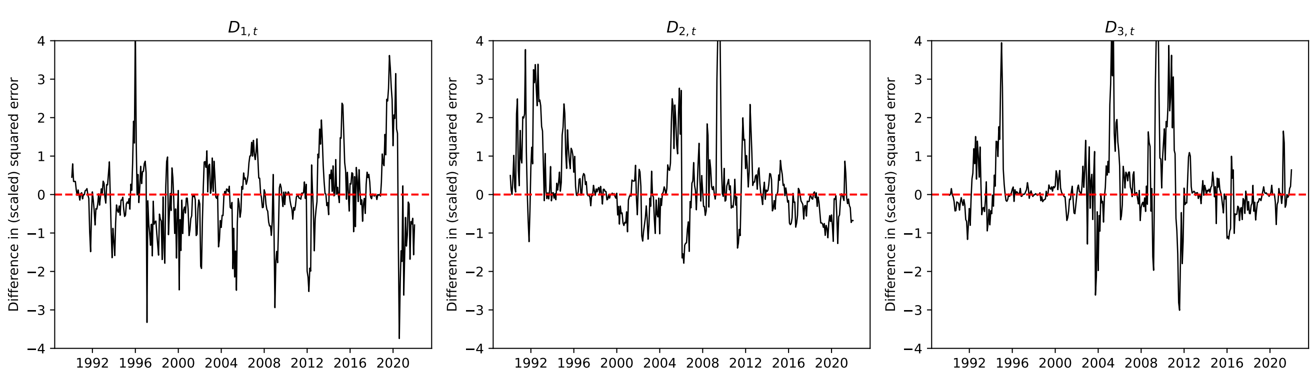


Figure B.7: This figure plots the evolution of the variables used in conditional predictive ability test. The out-of-sample period is 1990-2021.

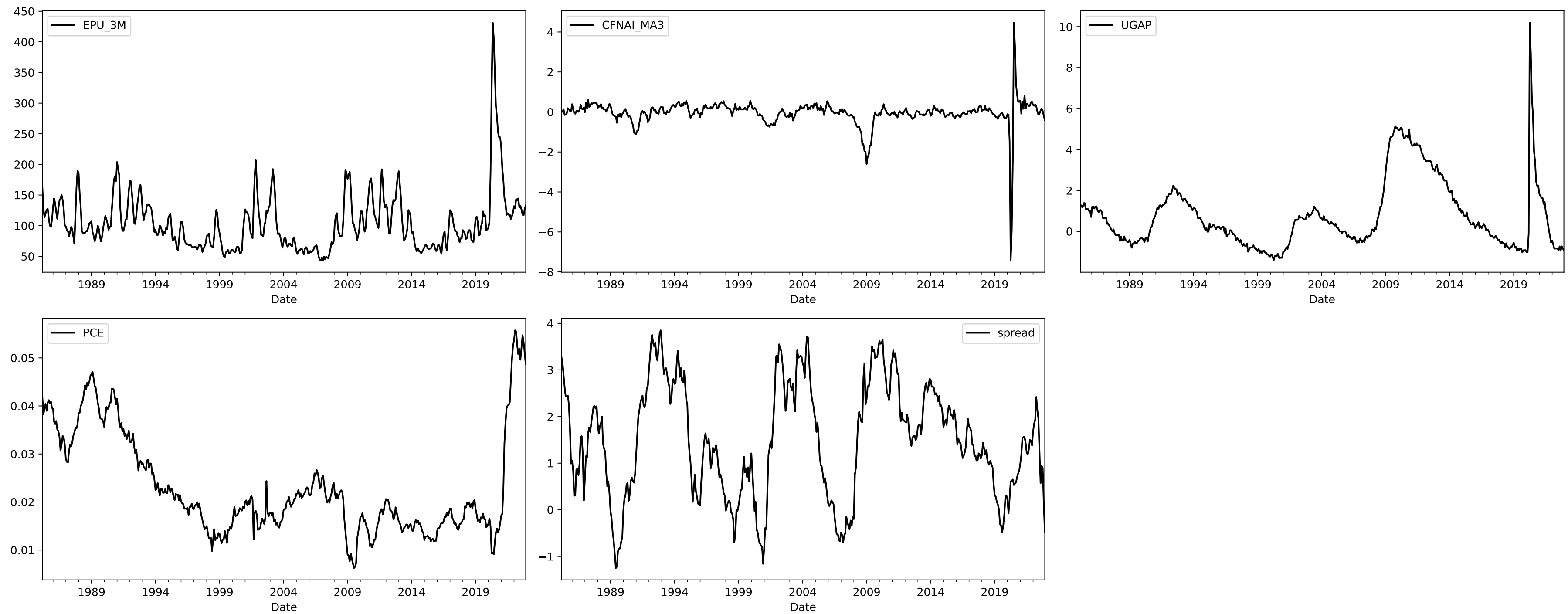


Figure B.8: Spectral decomposition of forward rates. The sample ranges from January 1973 to December 2021. We normalize the panel of forward rates and then compute the spectral decomposition of the associated correlation matrix. We display the cumulative variance explained by the eigenvalues.

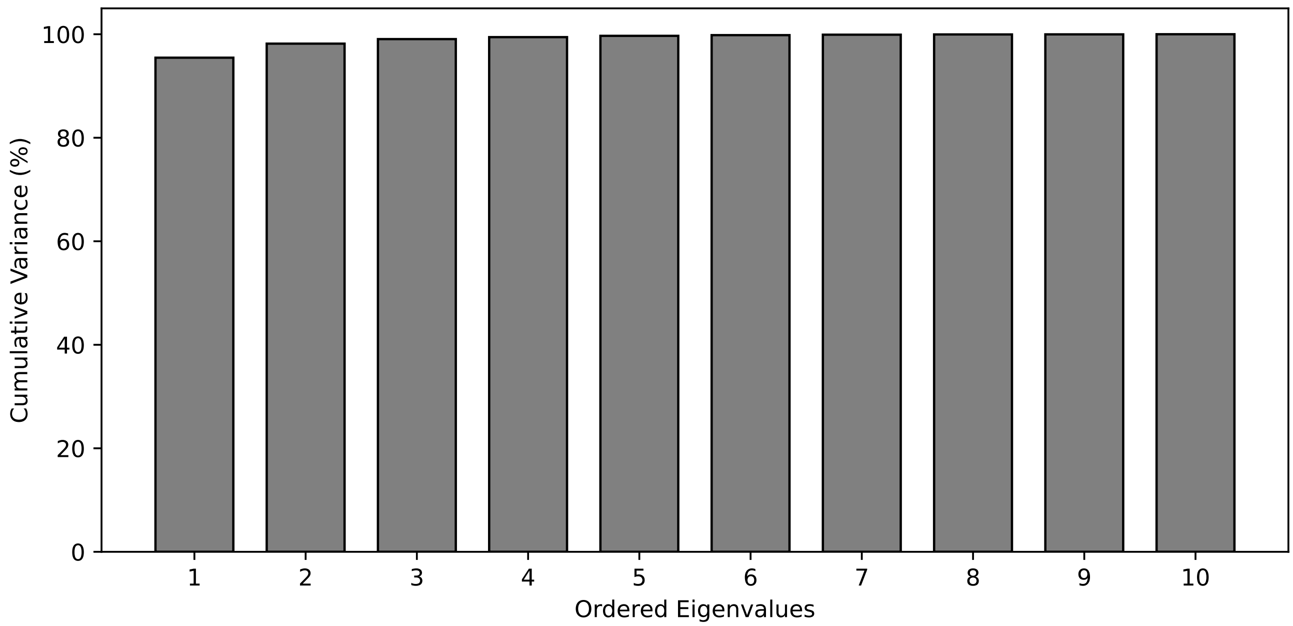
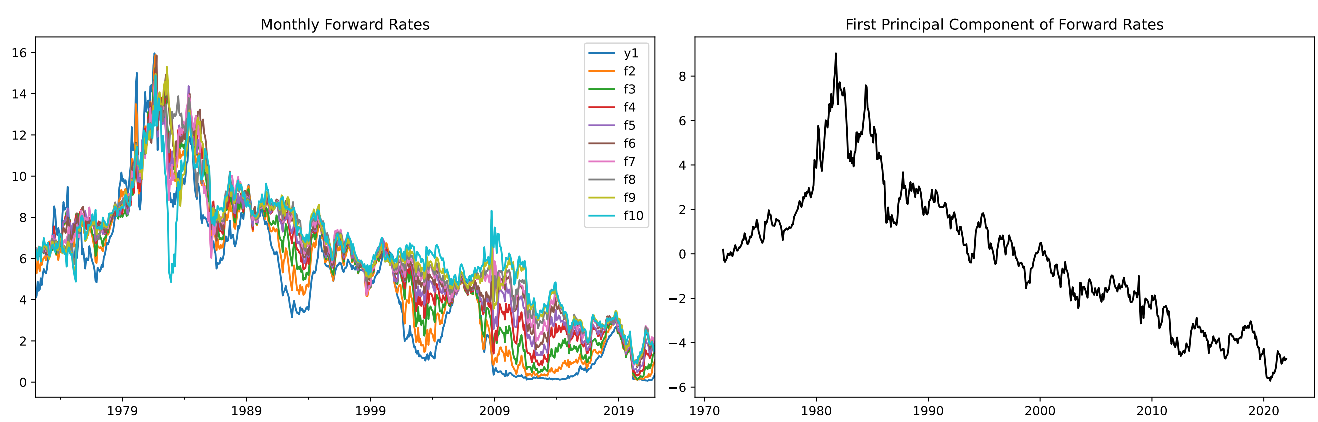


Figure B.9: Forward rates and the first principal component. The sample period ranges from January 1973 to December 2021.



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 setup 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 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.

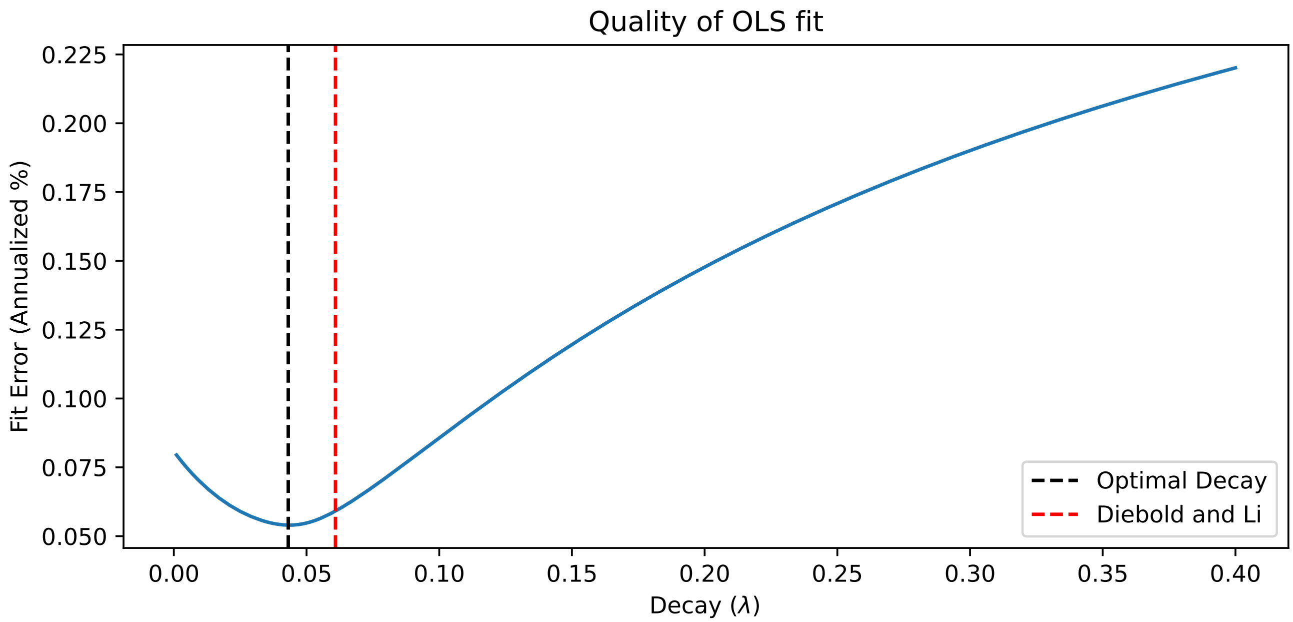
Figure C.2 assesses how big is the price in terms of in-sample fit that we are paying when using the same strategy as Diebold and Li (2006). It is not a high price. For each date, we compute the average squared residual in the cross-section of yields after fitting the model, take the square root, and plot it as a function of time. This time-series is a direct measure of how much information from the yield curve we lose by using a reduced-form model to summarize it. Until 2009, the three time-series are indistinguishable. Between 2010 and 2014 the performance of the Diebold and Li (2006) approach deteriorates 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-sample fit, however.[28] 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 in 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

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

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 of all yields from up to 120 months.



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 is only surpassed by the polynomial model during the heights of the Global Financial crisis in 2008 and a brief period between 2012 and 2014.

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.

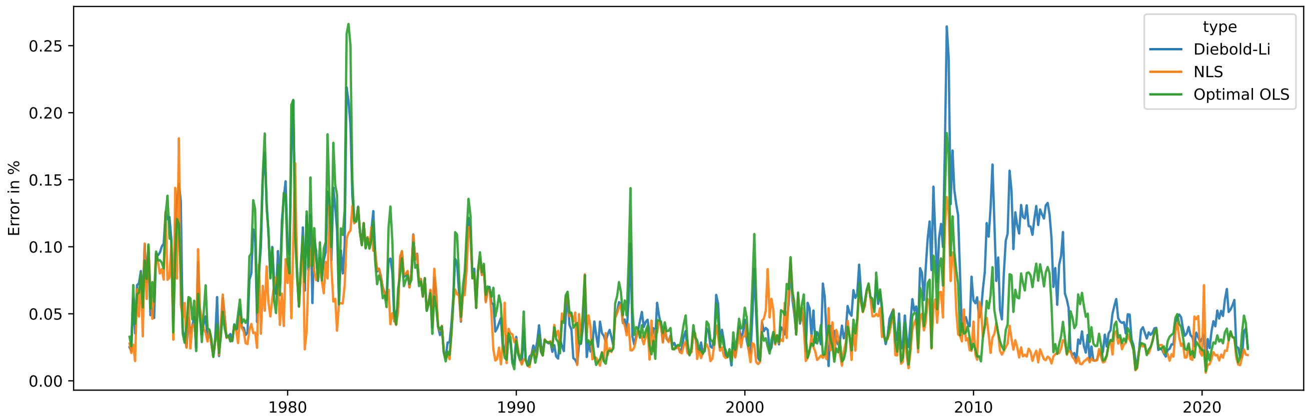
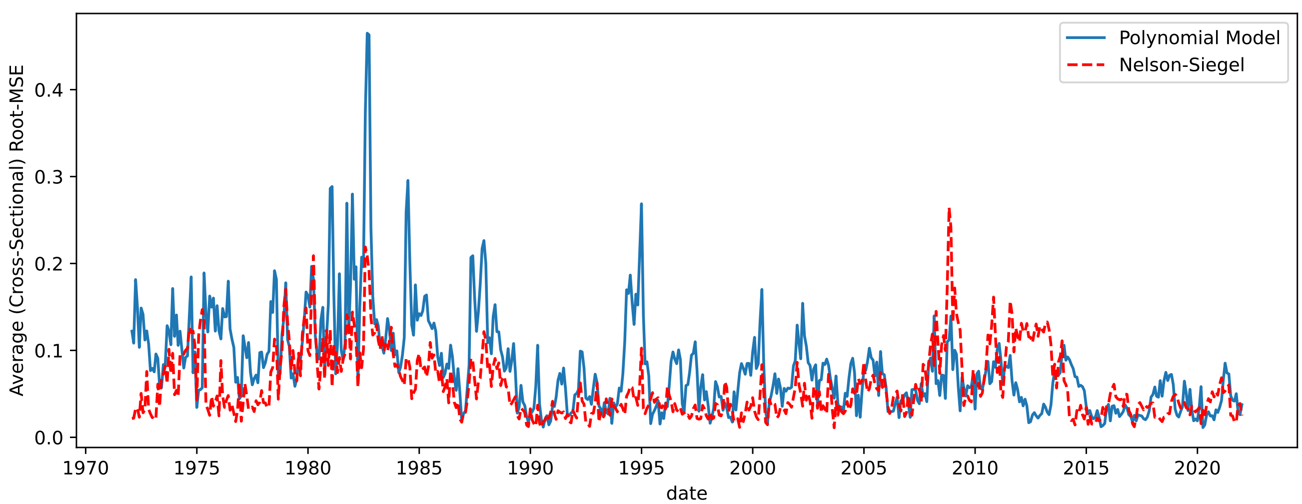


Figure C.3: This figure reports the same measure as Figure C.2 but comparing our Nelson-Siegel approach with the reduced-form model from equation (C.1). Both estimation procedures use all available yield from 1 to 120 months for each date.



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
AWHMAN	Labor Market	Avg Weekly Hours : Manufacturing	1
AWOTMAN	Labor Market	Avg Weekly Overtime Hours : Manufacturing	2

Continued on next page

FRED Code	Category	Description	Transformation Code
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
RPI	Output and Income	Real Personal Income	5
W875RX1	Output and Income	Real personal income ex transfer receipts	5

Continued on next page

FRED Code	Category	Description	Transformation Code
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 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 $[\beta_{2,t} + \beta_{3,t}]$

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

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.