Asymmetric Violations of the Spanning Hypothesis[*]

Gustavo Freire[†1] and Raul Riva[‡2]

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

First Version: March 17, 2023
This version: May 9, 2023

Preliminary draft: Please do not circulate

Abstract

There is a long debate over whether macroeconomic variables help predict bond returns after controlling for yield information. We document that violations of the spanning hypothesis are asymmetric across bond maturities: macroeconomic data is only useful for forecasting bond returns at shorter maturities. 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 macroeconomic variables only help predict the short- and medium-run factors that are relevant for short-maturity bonds. This predictability varies with the business cycle and monetary policy activity.

[*] We would like to thank Caio Almeida, Torben Andersen, Marcelo Fernandes, Robert Korajczyk, Costis Skiadas, Viktor Todorov, and Michel van der Wel for useful comments and suggestions.

[†] 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).

A natural question is then whether the yield curve itself contains all relevant information for predicting bond risk premia. As discussed by Bauer and Hamilton (2018), under certain assumptions 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] However, a number of papers show that macroeconomic variables have incremental predictive power for bond excess returns (Ludvigson and Ng, 2009; Joslin et al., 2014; Cieslak and Povala, 2015). While Bauer and Hamilton (2018) cast doubt on the validity of the inference from the in-sample regressions used in these papers, there is also recent out-of-sample evidence rejecting the spanning hypothesis (Bianchi et al., 2021).

In this paper, we document violations of the spanning hypothesis that are asymmetric across bond maturities. We provide in- and out-of-sample evidence that macroeconomic variables are useful for forecasting excess returns of bonds with shorter maturities (2- to 7-year), while for longer maturities the yield curve benchmark cannot be outperformed.[2] In fact, the predictive power of macroeconomic data relative to the yield curve decreases monotonically with the bond maturity. This has largely been ignored by the previous literature, which often focused on predicting an average of bond returns across maturities.

To help explain this pattern, we provide a new decomposition of bond returns. Assuming 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 the maturity, the higher the relative importance of the long-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 help predict both the short- and medium-run factors, while we cannot reject that the yield curve spans the relevant information for the long-run factor. This result holds both in- and out-of-sample. 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) methods. Across these different methods, the results are the same as for the principal components: macroeconomic variables only matter for predicting the short end of the yield curve.

The findings described so far are inherently unconditional in that they test whether there are violations of the

[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.

[2] We consider different sets of principal components factors extracted from a high-dimensional data set of macroeconomic variables (McCracken and Ng, 2016) as in Ludvigson and Ng (2009).

spanning hypothesis. 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 with the business cycle: macroeconomic variables afford incremental information during periods of higher economic activity indicated by higher inflation, lower unemployment and higher industrial production growth. On the other hand, macroeconomic data is relevant for forecasting the medium-run factor during periods of monetary easing and in special during the Zero Lower Bound. In other words, violations of the spanning hypothesis for the short end and middle of the yield curve depend on business cycle fluctuations and the current state of monetary policy.

We also leverage the Fluctuation test from Giacomini and Rossi (2010) and show that we cannot reject the much stronger null hypothesis of point-wise equal predictive ability of macroeconomic variables relative to yield information for forecasting the longer end of the yield curve. In contrast, rejections for the short and intermediate parts are abundant and robust to different forecasting window sizes. We also provide suggestive evidence that these violations become stronger at the intermediate part of the yield curve after 2008.

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 motivates our analysis by documenting asymmetric violations of the spanning hypothesis. Section 4 provides a 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 and macroeconomic information. Section 6 discusses the time-varying patterns of spanning hypothesis violations. Finally, Section 7 concludes the paper.

Related Literature

Our paper relates to the vast literature studying the term structure of interest rates. 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. 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) document that nonlinearities 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.

We contribute by providing new evidence that violations of the spanning hypothesis are asymmetric in the dimension of the bond maturity. In particular, macroeconomic variables help predict returns of bonds with shorter maturities. To help explain these patterns, we provide a new decomposition of bond excess returns using a dynamic version 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. Hännikäinen (2017) examines the predictive power of Nelson-Siegel factors for future industrial production. In contrast, we investigate whether macroeconomic variables help predict future realizations of the Nelson-Siegel factors after controlling for the current yield curve.

We show that macroeconomic data contain incremental information relative to the yield curve only for factors capturing behavior in the short end and middle of the yield curve. Relatedly, Ang and Piazzesi (2003) model the joint dynamics of bond yields and macroeconomic variables in a Vector Autoregression imposing no-arbitrage. Using variance decompositions, they find that macro factors primarily explain movements at the short and middle parts of the yield curve. While their results depend on a particular parametric specification, our findings are based on in- and out-of-sample predictive evidence. Importantly, we connect such patterns to the asymmetric bond return predictability across maturities afforded by macroeconomic data.

Dewachter and Lyrio (2006) estimate a reduced-form model for the joint dynamics of inflation expectations, a few macroeconomic variables and the term structure. They find that the level of the yield curve is related to long-run inflation expectations while the slope and curvature seem to be related to the business cycle and the stance of monetary policy. 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. They also find a relation between the business cycle and the slope factor. 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 dynamics observed on data and not necessarily in forecasting.

Finally, we find that asymmetric violations of the spanning hypothesis are linked to business cycle fluctuations and the current state of monetary policy. These links are hard to reconcile with canonical theory because the violations should not be present anywhere in the first place. Sarno et al. (2016) studies how bond return predictability is related to economic uncertainty while Gargano et al. (2019) studies holding bond return predictability over the business cycle. Their analyses focus on in-sample exercises, however. More closely related to us, Borup et al. (2023) provide evidence that violations of the expectations hypothesis are state dependent and can be predicted by economic activity also leveraging the conditional predictive ability framework from Giacomini and White (2006). One fundamental difference is that they do not concentrate on violations of spanning hypothesis, i.e., they consider a different benchmark. Additionally, to the best of our knowledge, we are the first to use a Nelson-Siegel approach to decompose the yield curve and analyze violations of the spanning hypothesis through these factors.

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).

Our choice of yield curve data set represents an improvement over other commonly used sources of yield curve data for the US, namely data sets constructed under either the methodology from Fama and Bliss (1987) or from 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 be contrasting 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. 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 chose 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. 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.[3] We start the sample in 1973 since the 10-year bond started being traded in late 1972.

[3] 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

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).[4] Table 17 in Appendix C 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 9, 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 different forecasting exercises whenever researchers need a standardized and freely available "data-rich" environment. An earlier version was 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) used it to forecast excess bond returns, while Medeiros et al. (2021) used it in an inflation forecasting exercise.[5] We follow the standard practice of the literature and use fully revised data.[6]

3 Motivating Framework

In order to motivate our decomposition of the yield curve with Nelson and Siegel (1987) factors, we start analyzing 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 also a standard framework to analyze the spanning hypothesis used, for example, by Ludvigson and Ng (2009), Joslin et al. (2014) and Bianchi et al. (2021).[7]

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.[8] We denote by $rx_{t+12}(n)$ 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 $t+12$:

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

As it is obvious from our notation, this return is only known at time $t+12$. However, 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 rx_{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 $rx_{t+12}(n)$ and, equivalently, $y_{t+12}^{(n-1)}$. This should hold regardless of the maturity n. Macroeconomic indicators are natural state

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 use maturities only up to 120 months. We have also estimated the Nelson and Siegel (1987) factors using all available data since our estimation method does *not* rely on a balanced panel and confirmed that this constraint does not change the time-series properties of the factors in any meaningful way.

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

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

[6] 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 had no material difference in results.

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

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

variables to test the spanning hypothesis since many models macroeconomists and financial economists use tie together aggregate variables and the dynamics of interest rates.

Since the number of macroeconomic variables we have is large, it is not feasible to simply add all of them to a linear model $rx_{t+12}(n)$ and estimate it by OLS, for example. Using only a few variables would 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.[9] We design both an in-sample and an out-of-sample forecasting exercise.

3.1 In-Sample Evidence

We 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 the yield curve. We study the in-sample predictive regression

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

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. The forward rate for maturity n at time t is defined as $f_t^{(n)} = n \cdot y_t^{(n)} - (n-1) \cdot y_t^{(n-1)}$. 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, a test of spanning hypothesis is a test of whether $\gamma_n = 0$. We compute standard errors using HAC variance estimators (Newey and West, 1987).

Table 1 presents results for the in-sample regression in (2) 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.

Although we can reject the null hypothesis at usual levels for the coefficient on the first principal component across maturities, we note another interesting pattern. 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 11 in the 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.

Taken at face value, this result suggests that macroeconomic variables can enhance the forecast of excess bonds returns in an asymmetric way: they are more helpful for the shorter end of the yield curve than for the longer end. This type of analysis, however, is subject to the criticism of Bauer and Hamilton (2018). They show, using simulation evidence, that the persistence of the regressors and the fact that we use overlapping returns may deem usual HAC-based inference unreliable. To guard against this criticism and to further investigate the apparent

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

Table 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

asymmetry in our results, we use a fully out-of-sample forecasting exercise that does not rely on the typical in-sample predictive regression previous literature has emphasized.

3.2 Out-of-sample Evidence

Now, we estimate a linear model as in (2) with an expanding window, keeping track of the one-year-ahead forecasts. We start our pure out-of-sample period in January, 1990 following Bianchi et al. (2021). For instance, the forecast for January, 1990 is made with all data available up to January, 1989. We take principal components of the macroeconomic variables available up to January, 1989 and use them as regressors for the forecasting exercise. After fitting the model with the available data, we use the estimated parameters to generate a forecast for the returns that will be realized in 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 of the FRED-MD data set. In total, we have 384 fully out-of-sample forecasts.

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:

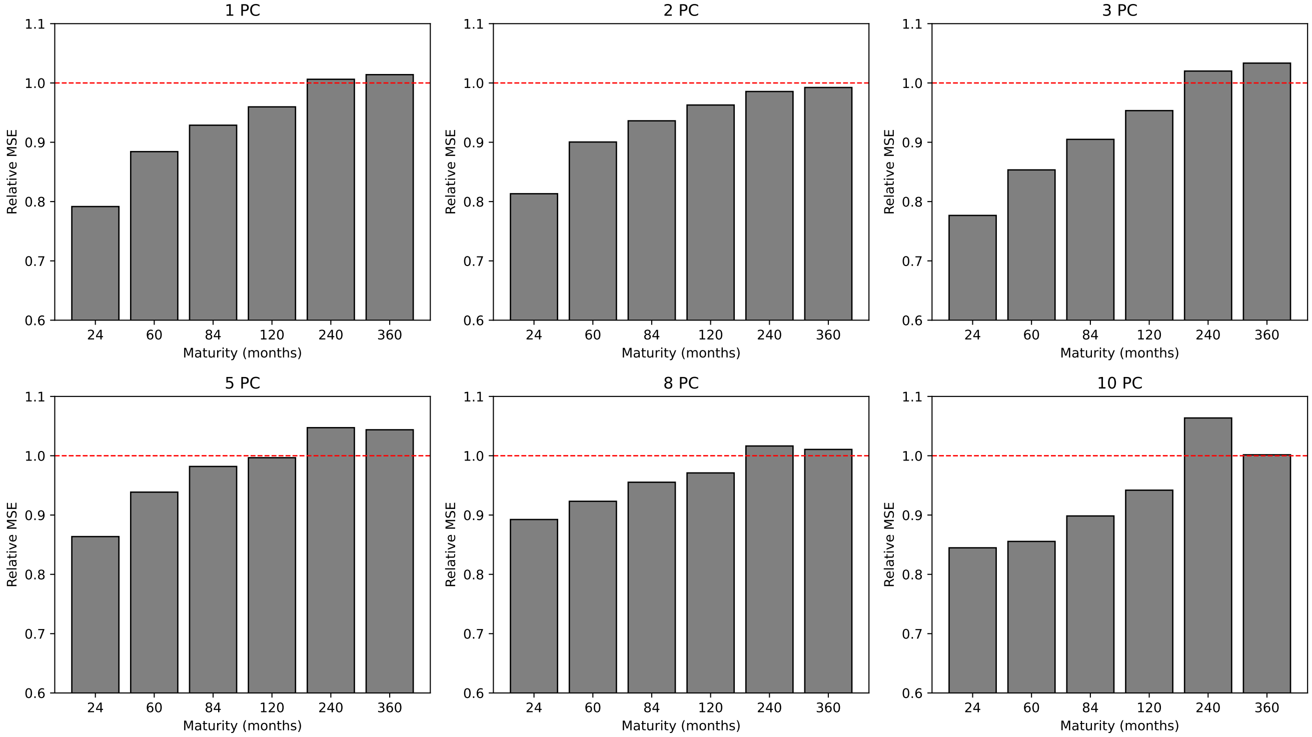
$$\text{MSE Ratio} = \frac{\sum_{t=t_0}^{T}(rx_t(n) - \widehat{rx}_t(n))^2}{\sum_{t=t_0}^{T}(rx_t(n) - \widehat{rx}_{t|\gamma_n=0}(n))^2} \tag{3}$$

where $\widehat{rx}_t(n)$ is a forecast made at $t-12$ from a model that allowed $\gamma_n \neq 0$ while $\widehat{rx}_{t|\gamma_n=0}(n)$ 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. The initial pattern hinted by Table 1 is even stronger in Figure 1. The shorter maturities display generally lower MSE ratios, which implies that conditioning on macroeconomic variables was more helpful to forecast excess bond returns at shorter maturities than at longer maturities. Under the spanning hypothesis, the bars from Figure 1 should oscillate around unity. However, for 24-month maturity, we document a decrease of about 20% of the mean squared error. As the maturity increases, the ratio approaches unity and goes slightly above in some cases. This evidence suggests once more that violations of the spanning hypothesis seem to be stronger at the shorter end of the curve. For instance, the reduction 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. Our other option for C_t, namely the extraction of principal components from the yield curve itself, might not be as reliable with a smaller amount of observations in the time-series as it was with the full sample. 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 we did with the macroeconomic data. Results are reported in Figure 10 in Appendix B. The same asymmetry arises but the absolute decrease in the MSE is larger.[10]

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.



[10] That 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.

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 2. The null hypothesis is equal predictive ability, which would be implied by the spanning hypothesis.

Table 2: p-values (Diebold and Mariano, 1995) for testing whether macro data enhances forecasting using forward rates as controls. See the discussion in the caption of Figure 1. Variances for the Diebold and Mariano (1995) test are computed using the HAC estimator of Newey and West (1987).

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

A few patterns appear from these p-values. First, at the 5% level, we reject the spanning hypothesis for the 2, 5 and 7-year maturities using one, two or three principal components from the macroeconomic data. This is an important finding by itself since it is not subject to the criticism from Bauer and Hamilton (2018), which concerns mainly the literature using in-sample predictive regressions. 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 Modeling the Yield Curve

Taken together, the results from the previous section point at an asymmetry in the violation of the spanning hypothesis since we had more success rejecting it for shorter maturities than for longer ones. We now seek to develop a methodology that will enable us to interpret and quantify this finding more thoroughly. The first step in our approach is modeling the entire yield curve with a parsimonious reduced-form model. 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

[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).

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 unknown 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) showed 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 but slightly different.

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, 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. The parameter λ controls how fast this decay happens. The interpretation of β_2 as the slope of the yield curve steams 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. The fact that the loading on β_3 has a hump-shaped format motivates calling it the curvature factor.

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

1. We have precise interpretations of the factors themselves, by construction.
2. Principal component analysis, when used as a way to identify factors in an approximate factor structure setting, suffers from an identification problem. Factors and loadings in that case are identified only up to a rotation. In principle, there is no *a priori* best possible rotation. The Nelson and Siegel (1987) approach solves this identification problem by assuming a parametric form of loadings.
3. The implied price at time 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 any sensible economic model. This does not need to be case if we use, for example, some splines-based methodologies.

Perhaps more importantly, the parametric form of the loadings in equation (4) delivers an interesting 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.

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

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*

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

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 innovation in the short-term factors 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's 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 proof of this proposition is not particularly insightful and is relegated to the appendix.

The decomposition from Proposition 1 indicates that the predictability of the excess bond returns must be tied to the predictability of the factor levels 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 shouldn't 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 adopted 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]

[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.

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 have picked $\lambda = 0.0609$ 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. Below, we show that this decay value fits the data well and is very close to what we define as the optimal value conditional on our data. Additionally, it facilitates comparisons with other studies that followed the same methodology.

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.

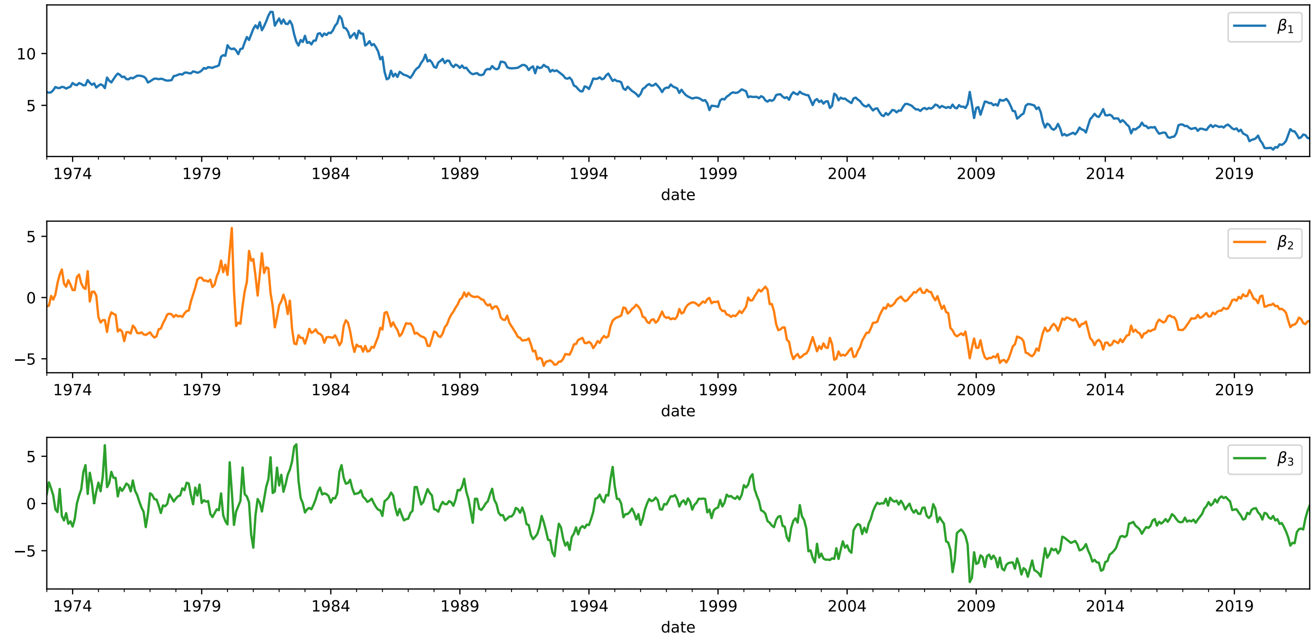


Figure 2 displays the estimated time-series for each factor. It is immediate to see that they are persistent time-series, which isn't a surprise since the the cross-section of zero-coupon yields is persistent as well. The scale of the long-run factor is also slightly greater than the realizations of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

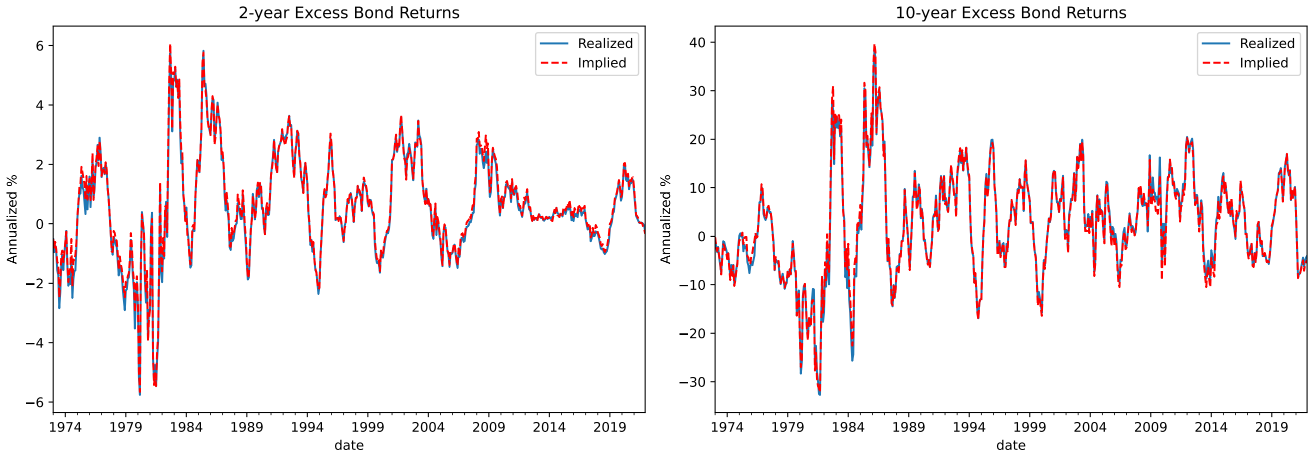
We have highlighted the advantages of the Nelson-Siegel approach but the question of whether the model is indeed able to fit the data is purely empirical. Importantly, given Proposition 1, a natural question is how well the excess bond returns implied by the dynamics of the Nelson-Siegel factors track the returns we observe from the realized zero-coupon yields.

Figure 3 is designed to answer this question. The blue solid represents the one-year excess bond returns computed using the observed zero-coupon yields, as defined in (1). The panel in the left shows returns for $n = 2$ years

Table 3: 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 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.



while the panel in 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).

4.2 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.

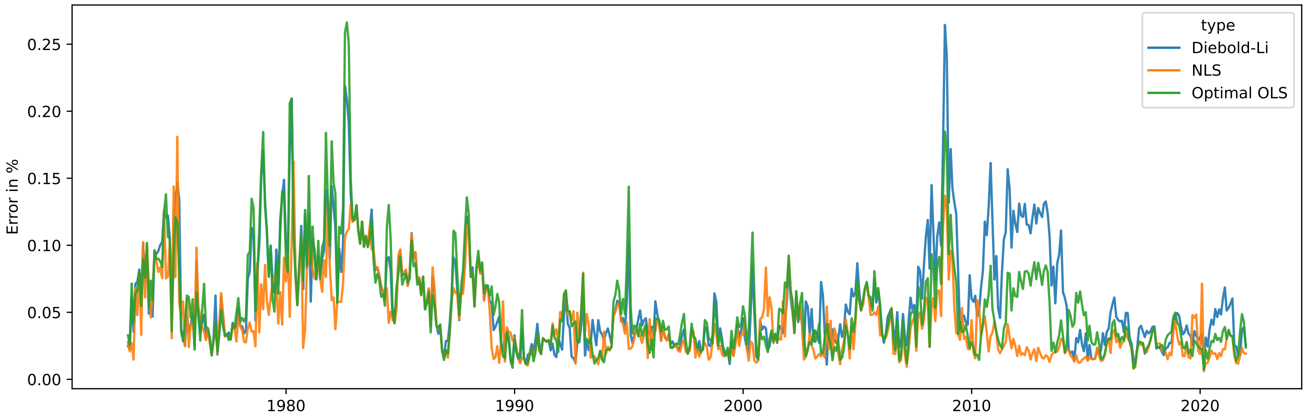
A second possible way to estimate factors and the decay parameter λ is using a two-step approach. For each given value of λ, we can estimate factors for all dates using the estimator in (7) and compute, for example, a time-series of the sum of squared residuals in the cross-section of yields. The average of this time-series can be understood as a measure of goodness-of-it 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 the appendix in Figure 11. 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 some 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 4 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.

Figure 4: 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.



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.[14] 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, 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.

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 don't find large gains by imposing no-arbitrage restrictions. Additionally, any gains seem less expressive after mid-1990s and they are concentrated in the very short end of the yield curve, like at one to three-month horizons. In general, the mean squared error decreases at most by 5% after the introduction of such restrictions.

5 Asymmetric 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. 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. We first analyze in-sample predictive regressions where the factors are the dependent variables and then consider fully out-of-sample forecasting exercises.

5.1 Predictive Regressions

Our first analysis considers in-sample linear predictive regressions

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

in which we let C_t stack variables that seek to span information contained in the yield curve at time t and PC_t stacks the realizations of principal components of the FRED-MD data set. Two natural choices for C_t are either the

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

forward rates or the lagged values of our factors. Hence, in the first case we have $C_t = \left(y_t^{(1)}, f_t^{(2)}, f_t^{(3)}, ..., f_t^{(10)}\right)'$ and in the second case we have $C_t = (\beta_{1,t}, \beta_{2,t}, \beta_{3,t})'$. None of our conclusions will be sensitive to this choice.

This exercise is similar to the our previous analysis of excess bond returns, with the difference that we have the factors as the dependent variables. We remark that our estimated factors are linear combinations of yields. Therefore, the spanning hypothesis would imply $\gamma_i = 0$ in the regression above. We compute standard errors using a HAC estimator based on Newey and West (1987). We also run separate regressions for each factor and different numbers of principal components in PC_t.

We report the estimates for γ_i in Table 4, where we use the forward rates to span the yield curve, and omit the other coefficients in the interest of space. We analyze four specifications for each factor, gradually increasing the number of principal components as we move from left to right. The last two rows display the adjusted R^2 for the baseline case where we include no principal components of the macroeconomic variables ("No Macro") and the limit case in which we add all possible variables ("All Macro"). The values are repeated for convenience. The sample period runs from 1973 to 2021.

Table 4: In-sample predictive regressions targeting the level of the factors as in (8). We only show estimates for γ_i. C_t stores the 1-year risk-free rate and the forward rates. 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") and the limit case when we include all macroeconomic variables ("All 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	-0.02	-0.04	-0.06	-0.03	-0.23***	-0.21***	-0.19***	-0.20***	-0.21***	-0.19***	-0.23***	-0.26***
	(0.03)	(0.03)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.05)	(0.04)	(0.05)	(0.06)	(0.06)
PC 2		0.04	0.05	0.02		-0.10**	-0.11**	-0.11**		-0.12**	-0.10*	-0.09
		(0.03)	(0.03)	(0.03)		(0.04)	(0.04)	(0.05)		(0.06)	(0.06)	(0.06)
PC 3		0.04	0.05*	0.04		0.08*	0.07	0.07		0.12**	0.15***	0.15***
		(0.03)	(0.03)	(0.03)		(0.04)	(0.05)	(0.04)		(0.05)	(0.05)	(0.05)
PC 4			-0.07**	-0.10***			0.01	0.01			0.05	0.07
			(0.03)	(0.03)			(0.04)	(0.05)			(0.04)	(0.05)
PC 5			0.13***	0.12***			-0.05	-0.05			0.11*	0.12**
			(0.03)	(0.03)			(0.05)	(0.04)			(0.06)	(0.06)
PC 6				-0.14***				0.06				0.09
				(0.05)				(0.07)				(0.07)
PC 7				0.01				-0.05				-0.12**
				(0.03)				(0.05)				(0.05)
PC 8				0.00				-0.18***				-0.19**
				(0.04)				(0.06)				(0.09)
R-squared Adj.	0.88	0.88	0.89	0.90	0.46	0.49	0.50	0.52	0.47	0.50	0.51	0.53
N	588	588	588	588	588	588	588	588	588	588	588	588
R-squared Adj. (No Macro)	0.88	0.88	0.88	0.88	0.34	0.34	0.34	0.34	0.42	0.42	0.42	0.42
R-squared Adj. (All Macro)	0.92	0.92	0.92	0.92	0.63	0.63	0.63	0.63	0.64	0.64	0.64	0.64

One measure of the usefulness of macroeconomic data when forecasting factors is the change in the R^2 from the baseline case to the R^2 attained by each specification. For the long-run factor, it barely increases, starting at 88% and moving to 90% in the most complete specification. In fact, even if we stretch the methodology to its limit and use all the information available, the R^2 will only increase two percentage points more. For both the short and medium factors, the difference is more pronounced. For β_2, the R^2 increases 18% (from 34% to 52%) when we compare the most complete specification to the baseline case. For β_3 we have an increase of eleven percentage points.

When we focus on the point estimates of γ_i, we find that we cannot reject the null hypothesis for at least the first three principal components in the case of β_1. On the other hand, we do reject the null hypothesis in the case of PC 1 and PC 2 for β_2 and PC 1 and PC 3 for β_3. This evidence is consistent with the analogous results for excess bonds returns from Table 1, even though we use a different set of variables to span the yield curve. The shorter end of the yield curve, here summarized by β_2 and to some extent by β_3, seems indeed more predictable when we allow for $\gamma_i \neq 0$. That is not the case for the longer end of the curve, summarized by β_1, whose predictability barely increased after we included principal components of the FRED-MD data set. Table 12, in the appendix, repeats this exercise but uses lagged values of the factors as baseline controls. Although we start rejecting the null of $\gamma_i = 0$ for PC 1 and PC 2 for β_1 as well, the difference in the adjusted R^2 from the baseline case to the most complete specification is even starker. It barely increases for β_1, while for β_2 it increases more than fivefold (from 7% to 40%).

5.2 Out-of-sample Evidence with Principal Components

Although the evidence based on the predictive regressions provides support to the conclusion that there is an asymmetry in the violation of the spanning hypothesis across the yield curve, this evidence is subject to the criticism of Bauer and Hamilton (2018). The dependent variables and the the regressors are fairly persistent and we work in an environment with overlapping observations. Based on the results in Bauer and Hamilton (2018), we entertain the hypothesis that our evidence so far might be in fact weaker than what the predictive regressions indicate. In order to bypass this concern, we design a fully out-of-sample forecasting exercise targeting the factors. We will analyze the prediction error with and without information from the macroeconomic variables. We emphasize that, under the spanning hypothesis, the macroeconomic variables shouldn't enhance the forecast of any factors, conditional on the information already spanned by the yield curve.

We deploy an expanding window forecasting exercise, as we did in Section 3.2. We use a linear forecasting model as in (8). 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. Hence, the time-series of PC_t does not contain any look-ahead biases. That would not be the case if we had used the entire sample to compute the principal components. 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. Notice that this measure can be negative if the forecast created by a given method is worse than the benchmark itself in terms of mean squared error. We choose to report this measure instead of a ratio of mean squared errors for two reasons. First, it provides a measure of absolute performance instead of just a relative one. Even when the addition of macroeconomic data enhances the forecast of a factor, it would also make sense to ask whether this forecast can beat a simple benchmark model. Showing that macroeconomic data can enhance the forecast of a factor is of little interest if this method cannot even beat a simple benchmark.

Two natural choices for benchmarks in our context are the historical mean and a random walk. If one believes the factors are stationary time series, reversion to the mean is expected and the historical mean is a consistent estimator for the unconditional one. Conversely, if one believes that factors have an unit root, a simple random walk becomes a sensible choice.

Additionally, since the factors are persistent series, we also target their innovations directly. This seeks to alleviate the concern that forecasting potentially integrated time series with (mainly) stationary 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

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

Table 5: OOS R^2 using the historical mean as benchmark. 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 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 linear models in (2) to make the forecasts. Principal components are extracted in a sequential way so they introduce no look-ahead bias. Panel A lets C_t store the three lagged Nelson-Siegel factors while Panel B controls for the forward rates. 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: Controlling for Lagged Betas

	R^2_{OOS}						p-values			
Target	No Macro	1 PC	5 PC	8 PC	10 PC	All Macro	1 PC	5 PC	8 PC	10 PC
Beta 1	0.87	0.91	0.90	0.91	0.90	0.72	0.01	0.03	0.01	0.04
Beta 2	-0.23	-0.07	0.25	0.25	0.25	-0.41	0.05	0.00	0.01	0.01
Beta 3	0.26	0.31	0.37	0.36	0.36	0.04	0.14	0.05	0.06	0.08
Innovation 1	0.90	0.92	0.92	0.92	0.92	0.81	0.06	0.02	0.04	0.13
Innovation 2	0.31	0.22	0.36	0.37	0.37	-0.40	1.00	0.22	0.20	0.17
Innovation 3	0.54	0.54	0.52	0.51	0.50	-0.16	0.51	0.72	0.77	0.87

Panel B: Controlling for Forward Rates

	R^2_{OOS}						p-values			
Target	No Macro	1 PC	5 PC	8 PC	10 PC	All Macro	1 PC	5 PC	8 PC	10 PC
Beta 1	0.91	0.91	0.92	0.92	0.92	0.78	0.57	0.36	0.13	0.18
Beta 2	0.21	0.25	0.41	0.40	0.41	-0.30	0.18	0.03	0.06	0.05
Beta 3	0.38	0.37	0.44	0.43	0.42	-0.03	0.68	0.11	0.17	0.25
Innovation 1	0.92	0.92	0.92	0.93	0.93	0.80	0.16	0.24	0.05	0.09
Innovation 2	0.28	0.30	0.37	0.35	0.37	-0.31	0.26	0.19	0.28	0.24
Innovation 3	0.52	0.52	0.54	0.54	0.53	-0.10	0.67	0.19	0.25	0.36

As before, we have at least two natural choices for C_t: the lagged factors or the forward rates. Table 5 displays results controlling for lagged betas in Panel A and controlling for the forward rates in Panel B. In both panels we use the historical mean 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 our Appendix A, Table 13 shows the analogous results but using the random walk as the benchmark. The p-values are the same because the benchmark itself is irrelevant for the test. 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 be only due to sampling noise. The columns labeled "All Macro" stretch the methodology to the limit and uses all data available.

We start focusing on Panel A. The R^2 for the level of β_1 increases from 87% to 91%. Although the Diebold-Mariano test deems that change statistically significant at the 5% level, the forecasting performance increase is quantitatively small. Additionally, the results from Table 13 imply that none of our specification could beat a simple random walk when forecasting the level of β_1 and controlling for the lagged Nelson-Siegel factor. In that sense, we don't consider this improvement in forecasting accuracy economically meaningful since it can't even a benchmark that is in fact spanned by the yield curve itself.[15]

The results for β_1 are dwarfed by the effect in β_2 and, to some extent, by the effect on β_3. For β_2, there is a 50 percentage point increase in the R^2 measure. It is negative in our baseline specification and reaches 25% with five or eight principal components. Importantly, the addition of macroeconomic variables makes us beat the random walk for β_2 ($R^2_{OOS} = 6\%$, from Table 13). For β_3 we have an eleven percentage point increase from the baseline specification to the one with five principal components. These changes are statistically significant at the usual levels. However, we can't beat the random walk again. This first piece of evidence shows that the addition of macroeconomic information in our forecasting design is more helpful when we try to forecast the short end of the curve than the longer end.

The analysis of the innovations as targets enable a similar conclusion although the overall effect if more muted. The increase in forecasting ability of β_2 measured by the out-of-sample R^2 is about six percentage points, against at most one percentage point for β_1 and no improvement at all for β_3. These changes are in general not statistically different than zero at the 1% level, however.

The column with all the possible information is included to showcase that there is a limit to this methodology in terms of how many principal components to include. If we were to keep increasing this number, performance would naturally decrease since there is a trade-off between spanning more information and also increasing estimation uncertainty. This highlights the challenge our out-of-sample methodology imposes: conditioning on more information might decrease the in-sample fitting error but this is far from obvious out-of-sample. Therefore, we see rejections of the spanning hypothesis through an out-of-sample improvement in forecasting performance as conservative approach. Additional information from macro variables need to be informative enough to make up for the increase in estimation uncertainty.

The analysis of Panel B reinforces our argument that there exists an asymmetry in the violation of the spanning hypothesis across maturities. First, no matter whether we directly target the level or the innovations in β_1, the increase in forecasting performance provided by macroeconomic data is marginal at best. It is never greater than one percentage point and not statistically different than zero at 5% in most cases. Moreover, Table 13 shows that we failed to beat the random walk when predicting the level of β_1 across different specifications.

Second, when we target the level of β_2, the out-of-sample R^2 increases almost twofold, from 21% to 41% if we include five principal components. This increase is statistically different than zero at the 5% level. Table 13 also shows that we beat the random walk by a large margin in that case. We also see a moderate improvement in the forecastability of β_3, whose R^2 increases from 38% to 44%, but this we can't reject this change is statistically different than zero.

The analysis of the innovations once more reveals that β_2 is the factor for which the addition of macroeconomic

[15] Notice that $\beta_{1,t}$ is just a linear combination of yields from time t. Hence, the random walk is a benchmark choice that would still be available under the spanning hypothesis.

data is most helpful. In contrast, the quantitative evidence of any improvements in β_1 is fairly weak. However it seems we have less overall less statistical power when forecasting the innovations in comparison to forecasting levels.

Taken together, our out-of-sample evidence points to a pattern that is largely consistent with the asymmetry from Figure 1. The macroeconomic information spanned by the principal components of the FRED-MD data set seems to be more helpful to forecast β_2 than the other factors, i.e., the violations of spanning hypothesis seem stronger at the shorter end of the yield curve.

5.3 Regularized Linear Models

Our evidence so far uses principal components of the macroeconomic variables to summarize the information spanned therein. Although simple to use, 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. But 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 documented, 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).

5.3.1 Notation

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). The full list of variables is Appendix C. 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

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

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

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

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

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

5.3.2 Forecasting Design

As we did in our out-of-sample exercise with principal components, we adopt an expanding window framework. It implies that for each forecast we make we have to rerun all the numerical optimization procedures, leading to a non-trivial computational cost.[19] Our first forecast is made for January 1990 and our last one is made for December 2021. If we start this exercise too early, we won't have enough data to perform successful numerical optimization

[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).

[19] We did all the implementation in Python, leveraging the classes defined in the `sklearn` library.

in the beginning of the sample. If we start it too late, we will not have a long enough time-series of errors to allow for a reliable assessment of the relative predictive ability. We compute 384 out-of-sample forecasts in total.

For any point in time, we divide the available data into two 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 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.

We highlight that this validation procedure is different than standard cross-validation or K-fold 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 designed to work with observations that are all assumed to be independent. It is, however, very similar to the approach adopted by Bianchi et al. (2021). We refer to Arlot and Celisse (2010) for an in-depth discussion of validation methods in different contexts. One limitation of this expanding window methodology is that data from the beginning of the sample will still impact estimation at the end of the sample. We will tackle this limitation later.

5.3.3 Forecasting Ability

We compare the out-of-sample R^2 with and without the macroeconomic variables across the three different models and different targets. Table 6 displays the results using the historical mean as the benchmark, whereas Table 14 in Appendix A displays results using the random walk as the benchmark. In Panel A, we let C_t stack the lagged betas and in panel B use the lagged forward rates 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 or without macroeconomic data.

We start by analyzing Panel A. When we try to predict the level of β_1, we see a slight increase in forecasting performance when we add macroeconomic variables (around six percentage points). But we note that the absolute values of the out-of-sample R^2 are much lower than we found in Table 5, where we used principal components. When we try to predict the level through the predicted innovation, we see the baseline performance increases and in fact matches both the performance with macroeconomic data and the performance from Table 5. There is a simple explanation for that. The long-run factor is the most persistent one and is close to being integrated. In a linear setting with principal components and no regularization, the estimated model would typically display a coefficient close to 1 for the lagged value of the long-run factor and smaller coefficients (in absolute terms) for other variables.[20] In a penalized regression context, however, using 1 as the coefficient on any given variable might be too costly since the loss function penalizes *both* the squared error and the norm of the coefficient being estimated. When we predict the innovation, on the other hand, we bypass this problem and essentially grant the method a free coefficient on the lagged value of the long-run factor.[21] We stress that macroeconomic variables did not improve the forecastability of the long-run factor at all for the innovations. Evidence from Table 14 also shows

[20] See, for example, the coefficients reported in Table 12.

[21] See the discussion in Hoogteijling et al. (2021) about the importance of stationarity/persistence in the prediction of bond risk-premia in a Machine Learning context.

Table 6: R^2_{OOS} from different models and for different targets, using the historical mean as the benchmark. 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 lets C_t store the three lagged Nelson-Siegel factors while Panel B controls for the forward rates. 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 Lagged Betas

Target	No Macro Data			All Macro Data			p-value		
	Ridge	Lasso	Elastic Net	Ridge	Lasso	Elastic Net	Ridge	Lasso	Elastic Net
Beta 1	0.59	0.59	0.56	0.65	0.66	0.63	0.03	0.00	0.00
Beta 2	-0.03	-0.01	-0.01	0.18	0.18	0.16	0.04	0.03	0.05
Beta 3	0.19	0.19	0.18	0.29	0.35	0.34	0.14	0.01	0.01
Innovation 1	0.93	0.94	0.94	0.91	0.94	0.94	0.99	0.59	0.48
Innovation 2	0.20	0.16	0.18	0.27	0.29	0.27	0.13	0.00	0.02
Innovation 3	0.58	0.57	0.57	0.48	0.58	0.58	0.95	0.36	0.26

Panel B: Conditioning on Lagged Forward Rates

Target	No Macro Data			All Macro Data			p-value		
	Ridge	Lasso	Elastic Net	Ridge	Lasso	Elastic Net	Ridge	Lasso	Elastic Net
Beta 1	0.63	0.61	0.60	0.66	0.65	0.65	0.03	0.01	0.00
Beta 2	0.05	0.03	0.02	0.17	0.21	0.19	0.14	0.02	0.03
Beta 3	0.29	0.26	0.26	0.37	0.37	0.38	0.04	0.01	0.01
Innovation 1	0.94	0.94	0.94	0.94	0.94	0.95	0.89	0.37	0.09
Innovation 2	0.05	0.03	0.01	0.18	0.25	0.23	0.05	0.00	0.00
Innovation 3	0.58	0.57	0.55	0.51	0.54	0.54	0.99	0.91	0.68

that we are incredibly far from the random walk when just targeting the level and we have the same forecasting ability when targeting the innovations. Hence, the addition of macroeconomic data did not help forecasting β_1.

In line with our previous evidence, we see non-trivial improvements in the forecasting performance for β_2 when we add macroeconomic variables. The out-of-sample R^2 leaves negative territory and increases up to 18% when we directly target the level of the factor. This increase is statistically significant at the usual levels. Although the relative increase is smaller when we predict the innovations, it is still around ten percentage points and statistically significant, with the exception of the Ridge method. We see this as further evidence on the violations of the spanning hypothesis at the shorter end of the yield curve. Unlike the long-run factor, we found that either targeting the level or the innovation leads to similar performance in the case of β_2. We also beat the random walk by a large margin (10% when targeting the level, 22% when targeting the innovation).

The evidence for the medium-run factor β_3 is mixed. When we try to forecast its level, we also see a non-trivial forecasting performance increase. For the Lasso and the Elastic Net, the out-of-sample R^2 increases almost twofold. These increases are statistically significant and also point towards violations of the spanning hypothesis in the intermediate part of the yield curve. Nonetheless, we don't see the same type of violations when we target the innovations. In fact, the performance slightly decreases after the introduction of macroeconomic variables in that case. And in neither case we are able to beat a random walk for β_3.

We now turn to Panel B, in which we let C_t contain the 1-year risk-free rate and the sequence of forward rates $\left[f_t^{(2)}, ..., f_t^{(10)}\right]$. The results are consistent to what we found when controlling for the lagged factors, which implies that they are not driven by our choices on how to span the information contained in the yield curve. For β_1, we see a slight increase in performance when we target the level and condition on macroeconomic variables. However, the performance in absolute terms is much worse than what we had with principal components. This is due to the interplay between regularization and a nearly integrated target discussed above. In contrast, when we target the innovation for the long-run factor, we see no sign of systematic improvement in forecastability due to the addition of macroeconomic variables to our estimation strategy. In this case, we have a similar level of performance in comparison to our results using principal components in absolute terms.

Panel B also shows that the improvement in the forecast of the short-run factor β_2 provided by the macroeconomic variables is large, i.e., a strong violation of the spanning hypothesis for the short end of the yield curve. The out-of-sample R^2 for the level of β_2 ranges from no more than 5% without any macroeconomic variables and increases up to 21% in the case of the Lasso. Aside from the case of Ridge, these changes are statistically significant at the 5% level. The results when we target the innovations in β_2 are also similar: the R^2 increases from 1% to 23% in the case of the Elastic Net, for example. We can also reject that these changes are statistically different from zero at the usual levels. No matter whether we look at the levels or at the innovations, we find strong evidence against the spanning hypothesis for the short end of the yield curve. We also highlight that we could only consistently beat a random walk, both for levels and innovations, for β_2.

The results for the medium-run factor in Panel B are also similar to what we found in Panel A. We see a moderate increase in forecastability by adding macroeconomic information when we target the level and this increase is significant at the 5% level. Instead, we do not see the same increase when we target the innovations. So the evidence regarding β_3 is still mixed. When compared to a random walk, the forecasts for β_3 are slightly worse than the benchmark when targeting innovations and quite worse when we target the level.

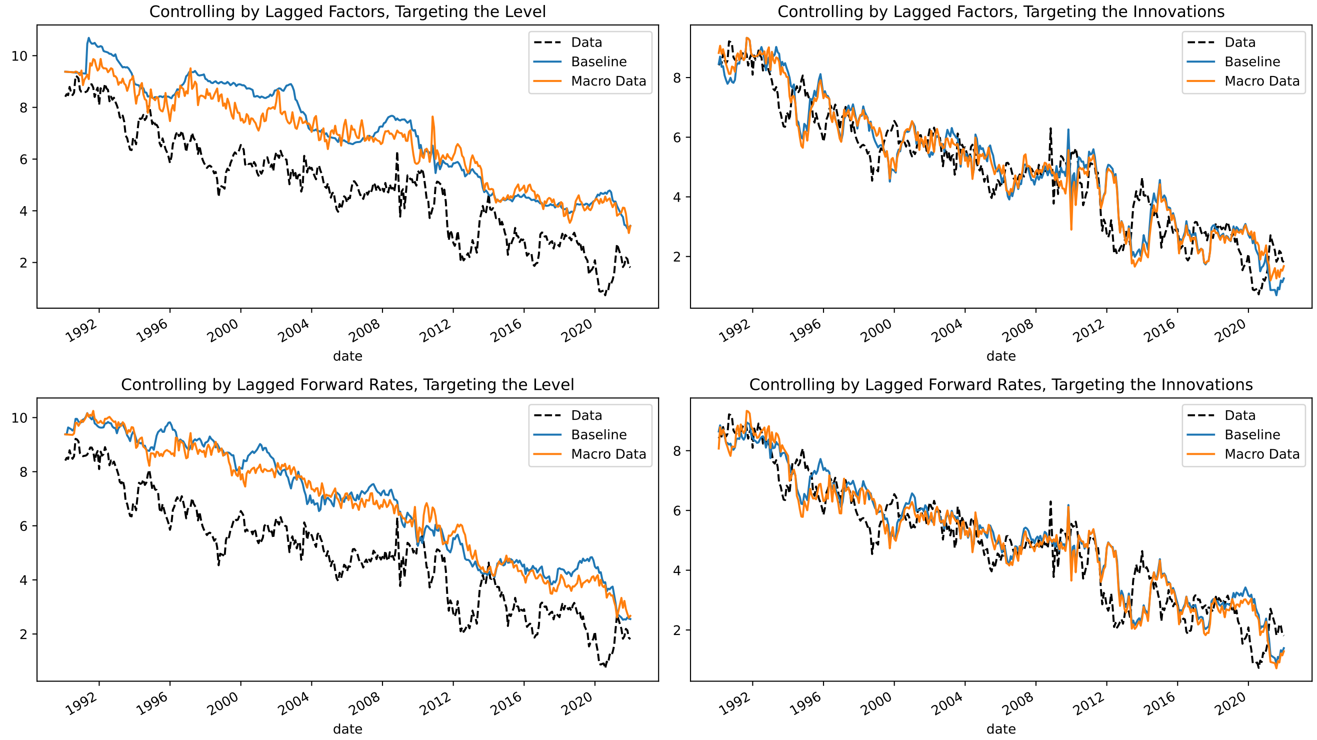
Finally, we use Figure 5 to further elaborate on our point regarding the interplay between regularization and a nearly integrated forecasting target. We display four panels. In each of them, the black dashed line shows the evolution of β_1 in our out-of-sample period. The blue lines display forecasts done without any macroeconomic variables, while the orange lines represent the forecasts produced using all information available in the FRED-MD data set. We focus on forecasts made by the Elastic Net since it is the most flexible method.[22] The panels on the first row span the yield curve using lagged value of the Nelson-Siegel factors whereas the ones on the second row the lagged forward rates. Panels on the left directly target the level of β_1 and the panels on the right target the innovations.

It is easy to see that 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 awful 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.

Therefore, we believe that the improvements provided by the macroeconomic variables when we forecast the level of β_1 with regularized models should be taken with skepticism, if not entirely disregarded. Although statistically different than zero, they are small and seem to happen in a context where we found the forecasting method itself

[22] Notice that our conclusions not depend on that particular choice. The forecasts created by these regularized models ended up being fairly similar across models. Extra results and forecasting plots are available upon request.

Figure 5: 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.



Controlling by Lagged Factors, Targeting the Level

Controlling by Lagged Factors, Targeting the Innovations

Controlling by Lagged Forward Rates, Targeting the Level

Controlling by Lagged Forward Rates, Targeting the Innovations

inappropriate for the target. We remark that this problem affects neither our targeting of the innovations nor the previous evidence with principal components since it's the interplay between persistence and regularization that cause the gap seen above. Since the other factors are not nearly as persistent as β_1, this issue does not arise in those cases.

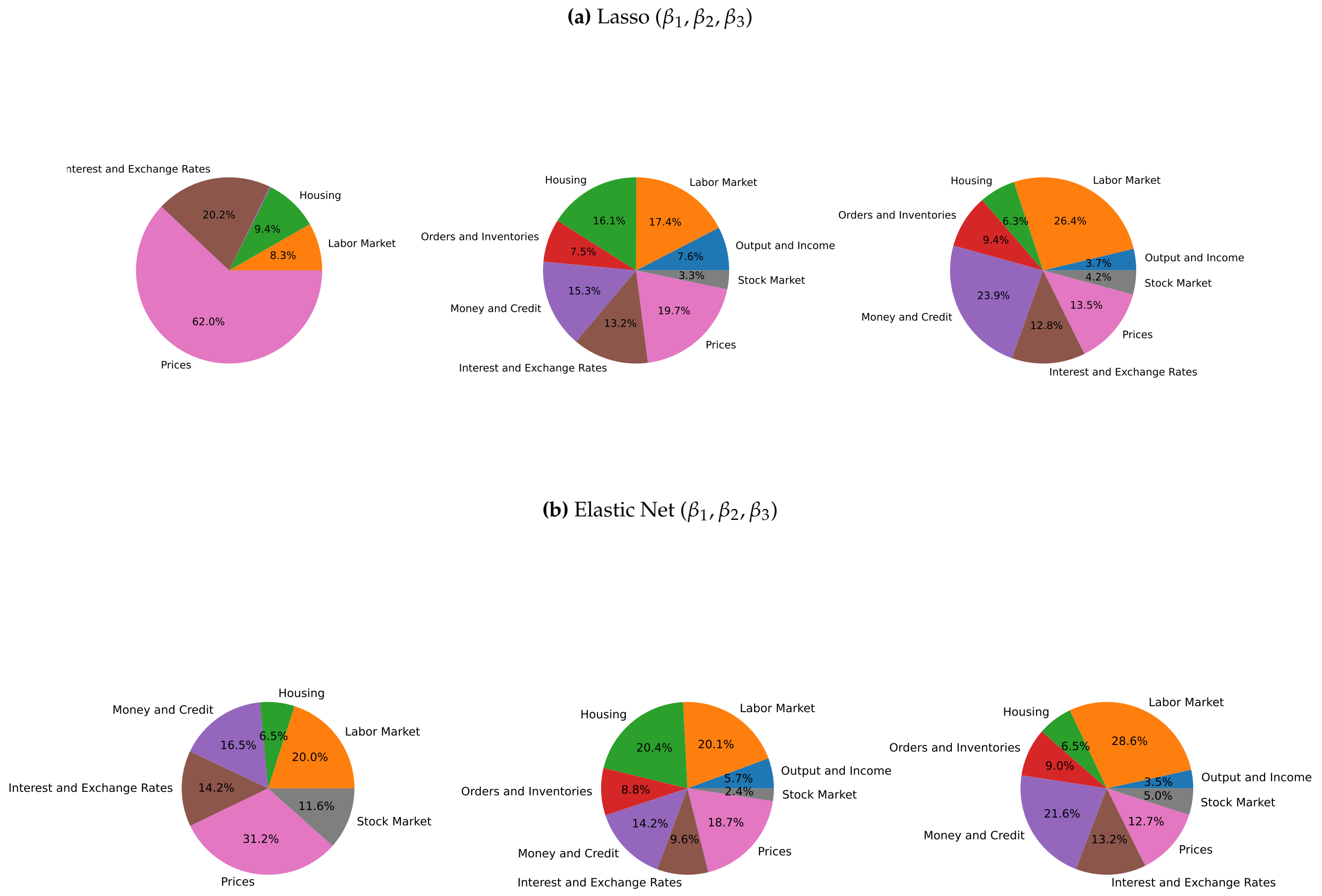
Taken together, the evidence from the regularized models agrees with the evidence from our approach using principal components. No matter how we span the yield curve or how we target β_2, we detect violations of the spanning hypothesis. On the other hand, violations through β_1 are much smaller and often statistically insignificant. The evidence for β_3 is mixed. In general, we found violations when targeting the level but didn't find the same result when analyzing the innovations. We also found that results were similar across the regularized models, with the Ridge method providing slightly worse forecasts than Lasso and Elastic Net when macroeconomic data is added.

5.3.4 Model Selection

As we mentioned above, Lasso and Elastic Net will 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 very 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.[23]

[23] See our Appendix C for a full list of variables and their classification.

Figure 6: 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 Nelson-Siegel factors. The out-of-sample period is January 1990 to December 2021.



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

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

We compute how frequently variables from each group were chosen and show results in Figure 6 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 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 pathology found when targeting its level, as shown in Figure 5. For the other two factors we focus on choices made when we target the levels since those were the cases in which macroeconomic data helped the most (see Table 6). We also focus on results when we control by the lagged values of the Nelson-Siegel factors.[24]

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

[24] Controlling by the forward rates leads to very similar pie charts which we omit for the sake of space but are available upon request.

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 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 and that leads to an improvement in forecasting is a manifestation of this lack of sparsity.

Our results for β_3 show that the methodology is also mixing a wide array of signals to predict the medium-run factor. But we note that the labor market indicators and the "Money and Credit" group, together, represent more than half of the total choices. The latter group includes measures of the monetary base and different monetary aggregates, aside from different spreads between corporate bond yields and the Fed Funds rate. We attempt no causal interpretation since our methodology is not designed for that, but we do note that Dewachter and Lyrio (2006) find that the yield curve curvature, closely linked to the medium-run factor β_3, is affected by variables related to monetary policy using a VAR approach. Their result is consistent with our findings. In fact, we will provide some evidence in the next section that violations of the spanning hypothesis through β_3 are apparently linked to the stance of monetary policy. Since monetary policy is also sensitive to labor market conditions, it makes sense to have the labor market indicators as another frequently chosen group of variables.

In summary, the evidence from the regularized models broadly agrees with the evidence from our linear regressions on principal components of macroeconomic variables. The violations of the spanning hypothesis are strong at the short end of the yield curve and weak at the long end.

6 Time-varying violations

We have emphasized out-of-sample evidence about the violations of the spanning hypothesis. Although we consider the forecasting setting we have better than pure in-sample predictive regressions the literature has commonly used, there are limitations in our methodology that we now try to overcome. First, the evidence so far depends on an expanding window forecasting design. This is useful because the amount of data available for estimation increases at each step. Nonetheless, this implicitly assumes some regularity in the underlying data-generating process. Data points from the beginning of the sample will affect the estimation step that will lead to an out-of-sample forecast. In this section we will work with a *rolling* window design and will reach the same conclusions as before.

Another limitation concerns the type of statistical test we have used to evaluate whether improvements in the forecastability of factors is statistically different than zero. Procedures like Diebold and Mariano (1995) and Clark and West (2007), for example, are *unconditional* tests. They can answer whether, on average, a model provides a better forecast than another one. We leverage the conditional predictive ability framework from Giacomini and White (2006) and the Fluctuation test from Giacomini and Rossi (2010) to understand not if but *when* the violations happen.

6.1 Evidence from Rolling Windows

We relax the assumption of an expanding window and now we work with a rolling window forecast design. Choosing the size of this rolling window implies a trade-off. A larger window provides more data for any estimation procedure we need to perform but leads to potentially more serially correlated forecast errors. On the other hand, a smaller window will imply noisy in-sample estimation which can lead to poor out-of-sample performance. We pick a window size of 180 months, but we have also tested 120 and 240 months and our results are not sensitive to that particular choice.

Given a window, we estimate linear model in (8) both with and without the principal components of the macroeconomic variables. We only use the data contained in each of these windows to extract the principal components. This ensures that have no look-ahead bias and that the extraction of the principal components in the later part of the sample is unrelated to data from the initial part. Our out-of-sample period starts in January, 1990 and ends in December, 2021 as before.

Table 7 reports the ratio between the mean squared error with and without information from the macroeconomic variables. A value below 1 implies that macroeconomic variables indeed help forecasting a given target. The spanning hypothesis implies that these values should oscillate around 1 due to sampling noise. We also report p-values for an associated Diebold and Mariano (1995) test where we assess whether any reduction was statistically significant.

Each of the different panels shows results for alternative choices for C_t. In the first and second panels we let C_t contain the lagged values of the Nelson-Siegel factors and the lagged forward rates, respectively. The third panel assumes that C_t has a single time-series, which is the first principal component of the forward rates. Cochrane and Piazzesi (2005) showed, with a different sample and a different source for the yield curve data, that the forward rates display a strong one-factor structure. We confirm this claim with a much more recent sample. Figure 12 in Appendix B displays the cumulative variance explained by the different eigenvalues from the spectral decomposition of $\left[y_t^{(1)}, f_t^{(2)}, ..., f_t^{(10)}\right]$. The first principal component explains more than 95% of the total variation. The panels in Figure 13 in Appendix B show the time-series for the forward rates and the first principal component. Along the lines of Cochrane and Piazzesi (2005), we take this as evidence that this first principal component summarizes well the information contained in the forward rates.

We start analyzing Panel A of Table 7. We first note that the inclusion of macroeconomic variables never helps the out-of-sample forecast of β_1 when we control for the lagged values of the Nelson-Siegel factors. In fact, the out-of-sample performance deteriorates. We stress that, as we increase the number of principal components from the macroeconomic variables, we potentially make more information available for the linear model but we also increase the number of parameters being estimated with the same amount of data. We pay the price of higher estimation uncertainty if we want to condition on more variables. For both β_2 and β_3 we see a nominal increase in forecasting performance but it's only significant at the 10% for β_3 with one principal component from the macroeconomic variables.

We find similar results when we control for the set of forward rates in Panel B. We see no signs of systematic

Table 7: Ratio of the MSE with and without principal components of the FRED-MD data set. We compute the ratio for each number of principal components and each factor. The last three columns report the p-value of the test from Diebold and Mariano (1995) where we test whether any reduction was statistically significant. The out-of-sample period ranges from January 1990 to December 2021. The first panels spans the yield curve through the lagged Nelson-Siegel factors. Panel B uses the lagged forward rates as control. Panel C uses only the first principal component of the forward rate.

Panel A: Conditioning on Lagged Betas

Number of Macro PCs	β_1	β_2	β_3	p-value (β_1)	p-value (β_2)	p-value (β_3)
0	1.00	1.00	1.00	-	-	-
1	1.02	0.95	0.91	0.67	0.37	0.06
2	1.08	0.98	0.91	0.94	0.46	0.10
3	1.28	0.96	1.12	0.99	0.41	0.93
4	1.23	0.95	1.14	0.92	0.39	0.89
5	1.29	1.05	1.25	0.93	0.59	0.96
6	1.31	1.07	1.40	0.92	0.60	0.96
7	1.31	1.07	1.41	0.92	0.60	0.96
8	1.16	1.14	1.53	0.85	0.67	0.97

Panel B: Controlling for All Forward Rates

Number of Macro PCs	β_1	β_2	β_3	p-value (β_1)	p-value (β_2)	p-value (β_3)
0	1.00	1.00	1.00	-	-	-
1	0.98	0.89	0.89	0.38	0.05	0.09
2	0.98	0.92	0.81	0.38	0.18	0.09
3	1.06	0.88	0.83	0.71	0.16	0.14
4	1.00	0.96	0.79	0.49	0.38	0.12
5	1.03	0.95	0.81	0.58	0.38	0.14
6	1.03	0.99	0.89	0.57	0.48	0.26
7	0.99	1.06	0.86	0.49	0.60	0.22
8	0.95	1.14	0.90	0.37	0.70	0.26

Panel C: Controlling for the Single Factor on Forward Rates

Number of Macro PCs	β_1	β_2	β_3	p-value (β_1)	p-value (β_2)	p-value (β_3)
0	1.00	1.00	1.00	-	-	-
1	0.98	0.86	0.80	0.36	0.10	0.00
2	0.99	0.73	0.70	0.50	0.01	0.00
3	1.20	0.67	0.75	0.97	0.01	0.00
4	1.39	0.82	0.81	0.93	0.23	0.03
5	1.44	0.88	0.85	0.93	0.34	0.08
6	1.42	0.82	0.88	0.93	0.26	0.18
7	1.40	0.82	0.89	0.92	0.26	0.20
8	1.25	0.84	0.93	0.88	0.30	0.31

improvements in the forecasting of β_1, which is the expected behavior under the spanning hypothesis. Conversely, we can reject the null for β_2 and β_3 at 5% with one principal component from the macroeconomic variables. Although there is still an improvement in the forecasting performance of β_2 and β_3 as we increase the number of principal components, it is not statistically different from zero. This hints at high estimation uncertainty. For example, to estimate the linear model from (8) with three principal components of macroeconomic data and all the forward rates, we have to estimate 14 parameters with 180 data points at each rolling window.

To circumvent the issue of estimation uncertainty, we in Panel C we control only for the first component of the forward rates. Figure 13 shows strong evidence that most of the information contained therein is well-summarized by this first principal component.[25] This approach brings the benefit of summarizing the information form the

[25] Cochrane and Piazzesi (2005) emphasize a similar point showing how a single factor they construct from forward rates has predicts excess bond returns, although they focus on an in-sample analysis that mixes together excess bond returns from different maturities.

forward rates with a single time-series, which decreases estimation uncertainty which leads to higher statistical power.

Panel C reports non-trivial forecasting gains for β_2 and β_3, which are statistically significant at the usual levels for 1, 2, and 3 principal components from macroeconomic data. For example, with three principal components taken out of the FRED-MD data set, the out-of-sample MSE for β_2 decreases by a third and by a fourth for β_3. These results reinforce the idea of stronger violations of the spanning hypothesis at the shorter end of the yield curve. On the other hand, we are yet to find systematic improvements in the forecasting of β_1. We find a small improvement for β_1 with one and two principal components but it is far from being statistically significant. Since the results controlling for the first principal component of forward rates displayed the strongest violations of the spanning hypothesis, we focus on them for the rest of the paper.[26]

6.2 Conditional Predictive Ability

Figure 7 displays the time-series for the squared prediction error when we control for the single factor from the forward rates. Each panel is dedicated to one of the Nelson-Siegel factors. In black, we show the out-of-sample prediction errors when no information from macroeconomic variables is included. The other three lines plot the analogous time-series when we control for one, two and three principal components of macroeconomic data, respectively. From now on, we focus on those forecasts since they are the ones for which we can reject the null hypothesis of equal predictive ability, as seen in Table 7.

First, we note that the lines are almost on top of each other for β_1, without a clear notion of what forecast is better. For β_2 and β_3 we see some similar pockets of predictability, i.e., when the black line is above the other ones. For β_2, that happens in the period 1992-1994 and also during 2002-2008. For β_3, the predictability seems stronger after 2009. This is consistent with Farmer et al. (2022) who show that predictability in equity markets arises in certain moments but disappears in others, creating these "pockets". The conditional predictive ability testing framework provided by Giacomini and White (2006) provides a natural environment to formally assess when the violations of the spanning hypothesis happen.

6.2.1 Notation

We let L_t^m denote the loss function associated to a forecast error of a certain variable x_t made by some model m. We adopt $L_t^m = (x_t - \widehat{x}_t)^2$ but this approach can accommodate different loss functions as well. The time-series in Figure 7 represent exactly L_t^m from different models. The approach from tests like Diebold and Mariano (1995) and Clark and West (2007) analyzes the following null hypothesis given two different models m and m':

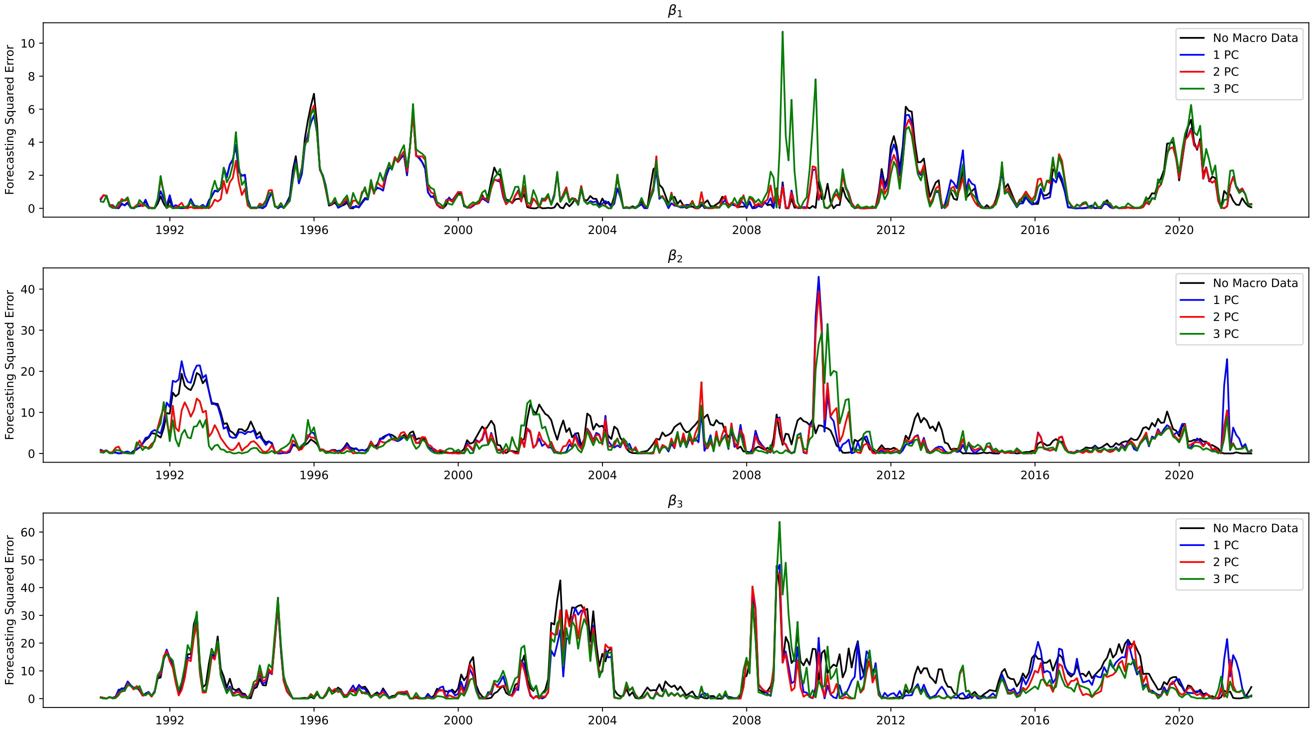
$$H_0: \quad \mathbb{E}\left[L_t^{m'} - L_t^m\right] = 0 \tag{14}$$

Under the null in (14), both models have the same unconditional predictive ability. A rejection of this type of null hypothesis implies that one model is on average a better forecaster than another one.[27] This approach is not informative, however, about when these differences in predictive ability will happen. It might well be the case that one model is better than the other on average but we cannot anticipate this gap in predictive performance.

[26] Our results are robust to the exclusion of the observations from November, 2008 to February, 2009 which is the period when the largest forecast errors happened. These coincided with the announcement of the first round of Quantitative Easing by the Federal Reserve. See Kuttner (2018) for a survey on the evidence of how these announcements moved the yield curve in an expected way.

[27] We emphasize that we have used the one-sided version of Diebold and Mariano (1995) since economic theory implies that addition of macroeconomic variables to our forecasting designs should not improve forecasting performance.

Figure 7: Time-series of squared out-of-sample prediction errors controlling by the first principal component of forward rates. For each factor, the black line displays the baseline forecast errors with no information from the macroeconomic variables. The other lines report forecast errors with different numbers of principal components from the FRED-MD dataset. The out-of-sample period ranges from January 1990 to December 2021. Forecasts are based on the linear model in (8). We take principal components in a sequential way to ensure no look-ahead bias.



Now we consider an information set $\mathcal{G}_t$ and postulate another null hypothesis:

$$H_0: \quad \mathbb{E}\left[L^{m'}_{t+h} - L^{m}_{t+h} \middle| \mathcal{G}_t\right] = 0, \quad t = t_0, ..., T-h \tag{15}$$

where h is a fixed forecasting horizon. We take $h = 12$. This null hypothesis implies that no $\mathcal{G}_t$-measurable function can predict the difference $L^{m'}_{t+h} - L^{m}_{t+h}$. From the point of view of time t, the differences in predictive ability should be indistinguishable from pure noise. For $h = 1$, for example, this is equivalent to saying that $\left\{L^{m'}_{t+1} - L^{m}_{t+1}, \mathcal{G}_t\right\}$ is a martingale difference sequence. It naturally nests (14) since we can always take $\mathcal{G}_t$ as the trivial sigma-field.

Giacomini and White (2006) propose a simple way to conduct the test. 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 $\mathbf{x}_t$. Let $\mathbf{z}_{t+h} \equiv \mathbf{x}_t \left(L^{m'}_{t+h} - L^{m}_{t+h}\right)$ and let

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

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

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

$$W \equiv T \cdot \mathbf{z}'_{t+h} \widehat{\Omega}_T^{-1} \mathbf{z}_{t+h} \xrightarrow{d} \chi^2_q \tag{16}$$

We note that this is an usual Wald-type statistic. We implement the test using a Bartlett kernel as in Newey and West (1987). The rejection of the null hypothesis at the confidence level α happens when $W > c_{q,1-\alpha}$ where $c_{q,1-\alpha}$ is the $(1-\alpha)$-quantile of a chi-squared distribution with q degrees of freedom.

In order to gauge the importance of different variables when the test is rejected, Giacomini and White (2006) suggest running the following predictive regression:

$$L^{m'}_{t+h} - L^{m}_{t+h} = \delta_0 + \delta' \mathbf{x}_t + u_{t+h} \tag{17}$$

The coefficients in δ are potentially useful to understand when the differences of predictive ability happen. Their test can also be interpreted as a joint test of $(\delta_0, \delta')' = \mathbf{0}$. In fact, when we include only a constant in (17) and test whether $\delta_0 = 0$, we recover the Diebold and Mariano (1995) test.[28] From now on, we adopt the convention that model m' includes only information spanned by the yield curve. Hence, $L^{m'}_{t+h} - L^{m}_{t+h} > 0$ implies that the addition of macroeconomic variables improved the forecasting performance (lower loss).

6.2.2 Running the Test

The econometrician has to choose the variables that go into $\mathbf{x}_t$. Canonical theory itself offers little guidance since, under the spanning hypothesis, there shouldn't be any sort of violations at all. Different empirical studies have linked violations of the *expectations* hypothesis to the macroeconomy (see for example Sarno et al. (2016), Gargano et al. (2019), Bianchi et al. (2021) and Borup et al. (2023)). Then it seems natural to include variables that are informative about the business cycle.

In related fashion, Cieslak (2018) documents large deviations between professional forecasters' expectations and realized yields that might induce the predictability we find. She shows that this effect is particularly strong when the Federal Reserve is easing interest rates. Therefore we also consider important introducing variables that are informative about current monetary policy in x_t.

In total, we consider the following variables, aside from a constant:

- "NBER": dummy variable that controls for recession periods dated by the National Bureau of Economic Research;
- "IP": monthly growth of industrial production;
- "Unemployment": monthly unemployment rate (U-3 measure from the St Louis Fred);
- "Inflation": rolling 12-month CPI inflation;
- "Fed Funds": effective Fed Funds rate at the last trading day of each month in our sample;
- "Monetary Policy Cycle": 12-month change in the effective Federal Fuds rate;
- "MP Activity": standard deviation of the changes in the Fed Funds rate over the last eight FOMC meetings;
- "ZLB": a dummy variable that controls for the Zero Lower Bound period;

[28] When we cannot reject the null hypothesis in (15), any evidence from the regression (17) should be considered with extra caution. Failing to reject the null hypothesis means that we cannot reject the possibility that variables in $\mathbf{x}_t$ are useless to predict the loss differential, i.e., predict violations. So one can make the case that (17) should be disregarded on those cases.

The first four variables are measures of the state of the business-cycle that are available at the monthly frequency, which is necessary for our tests. The last four variables measure the level of the policy rate, whether the Federal Reserve is in a tightening/easing cycle, how fast it is doing it, and a dummy variable to control for a period where the main policy tool was set to zero. Figure 14 in Appendix B displays the individual time-series of these variables.

Table 8 displays different specifications of (17) and the corresponding p-value for the conditional predictive ability test of Giacomini and White (2006) reported in the last row. The model we compare against the baseline is the one that uses two principal components of macroeconomic data, although we report the same table for the models with one and three principal components in Appendix A (see Tables 15 and 16). Each set of three columns corresponds to a different Nelson-Siegel factor. Standard errors for the individual coefficients are computed using Newey and West (1987).

Table 8: Auxiliary regressions for the conditional predictive ability test from Giacomini and White (2006), as described in (17). The last row displays the p-value of the test. We consider here the forecast made with two principal components of the FRED-MD data set. The dependent variable is the difference between the squared prediction error of the baseline model and the model with two principal components from the macro data. Standard errors are computed using Nelson and Siegel (1987). The out-of-sample period is January 1990 to December 2021. The first three columns concern the long-run factor, the three middle ones focus on the short-run factor and the last ones are dedicated to the medium-run factor. Stars denote significance at 10%, 5% and 1% respectively.

	ΔL_1 (1)	ΔL_1 (2)	ΔL_1 (3)	ΔL_2 (1)	ΔL_2 (2)	ΔL_2 (3)	ΔL_3 (1)	ΔL_3 (2)	ΔL_3 (3)
constant	-0.98*	0.19	-0.85	-0.64	0.04	-0.14	0.14	-0.06	0.75*
	(0.55)	(0.14)	(0.57)	(0.41)	(0.17)	(0.33)	(0.49)	(0.28)	(0.43)
NBER	-0.97***		-0.70**	-0.40		-0.94**	1.01**		0.40
	(0.24)		(0.32)	(0.55)		(0.46)	(0.41)		(0.26)
IP	-0.14		-0.13	0.18**		0.15	0.04		0.00
	(0.09)		(0.10)	(0.08)		(0.10)	(0.06)		(0.07)
Unemployment	0.11		0.13	-0.02		-0.13***	0.04		-0.16*
	(0.08)		(0.11)	(0.05)		(0.05)	(0.08)		(0.08)
Inflation	0.17**		0.24**	0.31**		0.49***	-0.19***		0.04
	(0.08)		(0.10)	(0.12)		(0.10)	(0.07)		(0.09)
Fed Funds		-0.03	-0.09**		0.03	-0.11***		-0.10*	-0.12**
		(0.03)	(0.04)		(0.05)	(0.04)		(0.05)	(0.05)
MP Cycle		0.11	0.09		-0.03	-0.17**		-0.10**	-0.12***
		(0.07)	(0.08)		(0.10)	(0.07)		(0.04)	(0.04)
MP Activity		-0.16	-0.18		-0.15	0.13		0.59*	0.59**
		(0.26)	(0.27)		(0.44)	(0.22)		(0.31)	(0.24)
ZLB		-0.04	-0.25		-0.35	0.17		0.54	0.92**
		(0.33)	(0.40)		(0.30)	(0.22)		(0.39)	(0.40)
N	384	384	384	384	384	384	384	384	384
R2 Adj.	0.11	0.04	0.15	0.22	0.03	0.30	0.13	0.22	0.27
GW p-value	0.21	0.57	0.17	0.01	0.02	0.06	0.02	0.02	0.09

We start noticing that we can never reject the null hypothesis for β_1 at usual confidence levels. Taken together with our previous evidence, this suggests that neither the business cycle nor the current monetary policy environment can predict the differential in predictive ability for β_1. This is consistent with our broader argument that the empirical evidence against violations of the spanning hypothesis for the longer end of the yield curve is weak.

Conversely, we reject the null hypothesis of equal conditional predictive ability at the 5% level for specifications 1 and 2 for β_2, and at the 10% level for the most complete specification. The coefficients from specification 1 imply that higher inflation and higher industrial production growth increase the loss differential, i.e., are related to stronger violations of the spanning hypothesis at the short end. However the coefficient on industrial production becomes insignificant after we control for more variables on the third specification.

In fact, our results from the most complete specification for β_2 are jointly consistent with both Borup et al. (2023) and Cieslak (2018). Borup et al. (2023) finds that violations of the *expectations* hypothesis are stronger in periods of higher economic activity. Likewise, we find that violations of the *spanning* hypothesis are stronger when the economy is not in a recession, unemployment is low and inflation is higher. Unlike them, we are able to show through our Nelson-Siegel decomposition that this happens at the shorter end of the yield curve and not across the entire spectrum of maturities.

Looking at monetary policy, we find that violations are stronger when the Federal Reserve is going through a monetary easing, which is represented by negative values of our Monetary Policy Cycle variable. Cieslak (2018) finds that these are moments in which there are large and persistent gaps between the expectations about yields from professional forecasters and the realized yields. Since the current yield curve reflects current expectations through prices, that can indeed lead to predictability from the point of view of the econometrician. One possible explanation for this predictability is that the information set used by a representative investor might be different from $\mathcal{G}_t$. But we emphasize that this interpretation is tentative and our methodology is not designed to rule out other explanations.

At the 10% confidence level we can also reject the null for β_3 for all three specifications considered. The results from the first column seem conflicting with our previous discussion since they point to a counter-cyclical pattern in the violations through β_3. But those coefficients are not statistically significant anymore after we control for the monetary policy block of variables. On the other hand, the evidence from the second and third specifications agree with each other: the violations through β_3 seem strongly predicted by monetary policy. They are stronger when the Federal Reserve is cutting rates and cutting them fast. This is represented by a negative coefficient on the monetary policy cycle variable and a positive one on monetary policy activity. These coefficients are statistically significant at the 5% level. Interestingly, we also find evidence that violations through β_3 were stronger during the Zero Lower Bound, which is line with Figure 7. The main "pocket" of predictability through β_3 comes from the period between 2009 and 2018, which includes the ZLB period.

Once more, this evidence is consistent with Cieslak (2018). It is also consistent with prior evidence from Dewachter and Lyrio (2006) who used an essentially affine term structure model to study the joint dynamics of latent yield curve factors and macroeconomic variables. They showed that these latent factors, here mirrored by $(\beta_1, \beta_2, \beta_3)$, are associated with long-term inflation expectations, current business cycle conditions and monetary policy, respectively. The puzzling aspect is why this information is not already spanned by the yield curve.

In Appendix A, Tables 15 and 16 reproduce the same analysis as in Table 8 but using one and three principal components to span the information from the forward rates. If we use one principal component, we are not able to reject the null hypothesis from Giacomini and White (2006) for any of the Nelson-Siegel factors across specifications. However, the individual coefficients from the regression in 17 tell the same story. Predictability seems stronger for β_2 and β_3 in times of higher economic activity and relaxation of monetary policy. With three principal components, a similar situation arises but we do reject the null hypothesis for β_2 and the all the individual coefficients agree with our previous discussion.

In summary, there are three main messages from this conditional predictive ability exercise. First, we find no evidence that we can predict when violations through β_1 will happen. As they don't even happen on average (or are very small), this is not a surprising result. But it brings extra support to our argument that the spanning hypothesis is less likely to be violated in the longer end of the yield curve. Conversely, we find that violations at the short end are stronger in periods of higher economic activity, summarized by lower unemployment and higher inflation. We also find that the relaxation of monetary policy contributes to these violations both through β_2 and β_3.

6.3 Predictive Ability Over Time

Both the unconditional and the conditional tests we have studied assume some stability from the data generating process. They use information from the whole series of out-of-sample forecasts to compare the relative predictive ability of different models. We now focus on the framework developed by Giacomini and Rossi (2010) which is designed to study how relative predictive ability evolves over time.

The null hypothesis we now entertain imposes in a much stronger version of (14) given two forecasting models m and m':

$$H_0: \quad \mathbb{E}\left[L_t^{m'} - L_t^m\right] = 0, \quad \forall t \in \{t_0, t_0+1, ..., T-1, T\} \tag{18}$$

where $\{t_0, t_0 + 1..., T-1, T\}$ is the set of dates for which forecasts were made. This null hypothesis implies not only that the models have unconditional predictive ability but also that they have point-wise equal predictive ability.

When comparing two different forecasting methods, one can imagine a situation in which one model is better at the beginning of the sample but the other one has the lead at the end. On average they might generate approximately the same forecasting errors but their behavior is certainly different point-wise. The Diebold and Mariano (1995) test has low power against this alternative while the framework developed in Giacomini and Rossi (2010) was designed to exactly tackle this question. Intuitively, the null hypothesis in (14) concerns the first moment of $L_t^{m'} - L_t^m$ while (18) uses information about the whole *path* of $L_t^{m'} - L_t^m$.

6.3.1 Notation

The test uses an auxiliary statistic that measures the local differential performance of models. First, the econometrician picks a window size of K observations and computes a normalized rolling mean of $L_t^{m'} - L_t^m$ over this window that we denote as F_t. We let $\mu \equiv K/(\text{total number of out-of-sample forecasts})$. We have the following formulation:

$$F_{t,\mu} \equiv \frac{1}{\sqrt{K} \cdot \widehat{\sigma}} \cdot \sum_{s \in \mathcal{I}} \left(L_s^{m'} - L_s^m\right), \quad \mathcal{I} = (t-K, t-K+1, ..., t) \tag{19}$$

where $\widehat{\sigma}^2 \xrightarrow{p} \lim\limits_{T\to\infty} \mathbb{E}\left[T^{-1/2} \sum\limits_{t=t_0}^{T} \left(L_t^{m'} - L_t^m\right)\right]^2$ is an estimator of the "scale" of $L_t^{m'} - L_t^m$.

Giacomini and Rossi (2010) use a Functional Central Limit Theorem to show that $F_{t,\mu}$ has an asymptotic distribution related to a functional of the standard Brownian Motion. Since the spanning hypothesis implies that the addition of macroeconomic variables should not make the forecasts more precise, we adopt the one-sided version of their framework, which uses the following null hypothesis:

$$H_0: \quad \mathbb{E}\left[L_t^{m'} - L_t^m\right] \geq 0, \quad \forall t \in \{t_0, t_0+1, ..., T-1, T\} \tag{20}$$

We reject this null hypothesis, for a given value of μ, whenever $\max_t F_{t,\mu} > \kappa_\alpha$ where κ_α is the $(1-\alpha)$-quantile of a non-standard distribution tabulated by their paper. Intuitively, we reject this null hypothesis whenever the model that conditions only on the spanning hypothesis becomes much worse than the other model considered (so it displays higher realizations of the loss function).

6.3.2 Running the Test

In order to conduct the test, we need to specify the constant μ. We consider nine values for μ, ranging from 0.1 to 0.9. In total, we have 384 out-of-sample forecasts. Table 9 displays window sizes and respective values for μ

Table 9: Test statistic and critical values for the one-sided test from Giacomini and Rossi (2010). For each factor, we compute the test static and compute its realization to the critical values for different window lengths parametrized by μ. In bold, we highlight the test statistic when the test rejects the null hypothesis at the 5% level. The out-of-sample period is from January 1990 to December 2021. The baseline forecast is always the one implied the spanning hypothesis, i.e., no added macroeconomic data. In the first panel, we add the first principal component of the FRED-MD data set. Panel B considers the case of two principal components and Panel C considers three principal components.

Panel A: GR Test with One PC

Window Size	μ	ΔL_1	ΔL_2	ΔL_3	Critical Value at 5%	Critical Value at 10%
38	0.1	1.698	2.602	3.111	3.176	2.928
76	0.2	1.526	**3.059**	**3.238**	2.938	2.676
115	0.3	1.561	**2.988**	**3.251**	2.770	2.482
153	0.4	0.779	**2.695**	**3.304**	2.624	2.334
192	0.5	1.056	2.437	**3.443**	2.475	2.168
230	0.6	0.778	2.104	**3.694**	2.352	2.030
268	0.7	0.663	2.095	**3.478**	2.248	1.904
307	0.8	0.827	1.954	**3.278**	2.080	1.740
345	0.9	0.852	1.857	**3.270**	1.975	1.600

Panel B: GR Test with Two Macro PCs

Window Size	μ	ΔL_1	ΔL_2	ΔL_3	Critical Value at 5%	Critical Value at 10%
38	0.1	2.332	3.142	**3.711**	3.176	2.928
76	0.2	1.266	2.390	**3.512**	2.938	2.676
115	0.3	1.032	2.400	**4.488**	2.770	2.482
153	0.4	0.797	**2.915**	**4.244**	2.624	2.334
192	0.5	0.453	**3.083**	**3.961**	2.475	2.168
230	0.6	0.174	**3.017**	**4.067**	2.352	2.030
268	0.7	0.667	**2.426**	**3.895**	2.248	1.904
307	0.8	0.495	**2.216**	**3.693**	2.080	1.740
345	0.9	0.425	**2.429**	**3.687**	1.975	1.600

Panel C: GR Test with Three Macro PCs

Window Size	μ	ΔL_1	ΔL_2	ΔL_3	Critical Value at 5%	Critical Values 10%
38	0.1	1.469	**3.514**	**3.221**	3.176	2.928
76	0.2	0.834	2.477	**3.437**	2.938	2.676
115	0.3	0.478	2.346	**4.105**	2.770	2.482
153	0.4	-0.448	2.605	**3.866**	2.624	2.334
192	0.5	-0.449	**2.746**	**3.383**	2.475	2.168
230	0.6	-1.280	**2.841**	**3.624**	2.352	2.030
268	0.7	-1.259	2.121	**3.427**	2.248	1.904
307	0.8	-1.268	1.936	**3.295**	2.080	1.740
345	0.9	-1.378	**2.120**	**3.090**	1.975	1.600

in the first two columns. The third, fourth and fifth columns displays the test statistic $\max_t F_{t,\mu}$ for each of the Nelson-Siegel factors. The two last columns displays critical values for the test both at 5% and at 10% levels. We make the values for the test static bold whenever it's significant at the 5% level. The table dedicates each of its three panels to different forecasting the models. From Panels A to C, we condition the forecasts on macroeconomic information using 1, 2 and 3 principal components, respectively.

The first clear pattern we highlight is that in no case we are able to reject the null hypothesis for β_1. No value of μ across the three panels delivered a rejection at least at the 10% confidence level. This is even stronger evidence in favor of the spanning hypothesis for the longer end of the yield curve. The point-wise null hypothesis we consider imposes very strong structure on the path of $L_t^{m'} - L_t^m$ and the data is not able to reject it. We interpret this result as a strong additional piece of support to our earlier evidence.

Symmetrically, rejections for β_2 and β_3 are abundant. For β_2, we reject the null at the 10% level almost always, with the exceptions of $\mu = 0.1$ in the first panel and $\mu \in \{0.2, 0.3\}$ in the last one. For β_3 we always reject at the 10% level and only fail to reject at the 5% level for $\mu = 0.1$ in Panel A. These results are also consistent with our previous evidence that uncovered stronger violations of the spanning hypothesis through β_2 and β_3.

6.3.3 Evolution of Local Relative Predictive Ability

Figure 8 displays the evolution of $F_{t,\mu}$ for $\mu = 0.4$ over time. Positive values mean that the macroeconomic variables enhanced the forecast of the Nelson-Siegel factors, at least on a local sense. More extreme values of F_t imply more extreme violations of the spanning hypothesis. Each of the three rows correspond to a panel in Table 9 while each of the columns correspond to a different Nelson-Siegel factor. The red dashed line depicts the 10% critical value. For β_1, we see that the measure proposed by Giacomini and Rossi (2010) fluctuates around zero with no clear trend and is always far from the critical value threshold. This agrees with our previous evidence of no systematic violations of spanning hypothesis through β_1.

We see the local measure F_t cross the threshold for both β_2 and β_3, regardless of whether we include one, two or three principal components from the FRED-MD data set. But their behavior follow opposite patterns. For the short-run factor β_2, the moments of stronger violations of the spanning hypothesis happened in the beginning of the sample. that explains why the local measure declines over time, in special after 2008. Alternatively, the local measure for β_3 is generally increasing, meaning the violations of the spanning hypothesis have become stronger over time. Remarkably, the evidence from the two first panels for β_3 also suggest 2008 as a point from which the violations got stronger. In Appendix B, Figure 15 reproduces Figure 8 using $\mu = 0.5$ and the visual evidence is unchanged.[29]

Since monetary policy acquired new aspects after 2008, namely unconventional measures from the Federal Reserve and the attainment of the Zero Lower Bound by the Fed Funds rate, it is intriguing that both β_2 and β_3 display the behavior seen in Figure 8. To formally test whether violations were any different after 2008, we conduct another exercise leveraging the testing framework from Giacomini and White (2006). Now we include in $\mathbf{x}_t$ two dummy variables. The first is 1 after 2008 and 0 otherwise. The second dummy variable is 1 during the period of the Zero Lower Bound and zero elsewhere.

Results for individual regressions, together with the p-values for the conditional predictive ability test, are displayed in Table 10. We analyze three different specifications, controlling for each of the dummies and then including both of them in the regression. We show three different panels as we perform the analysis on the forecasts that condition on 1, 2 and 3 principal components from the macroeconomic variables. First, we highlight we cannot reject the null hypothesis of equal conditional predictive ability across any of the panels for β_1. That agrees with the time-series for $F_{t,\mu}$ displayed in Figure 8 for the long-run factor.

We reject the null hypothesis at the 5% level for β_2 in both Panels B and C. The individual coefficient on "After 2008" is not significant after we also control for the ZLB dummy, however. That implies we do not have enough statistical power to infer that violations through β_2 became less severe after 2008. The results for β_3 are stronger, nonetheless. We never fail to reject the null hypothesis at the 5% level in any of the specifications and panels

[29] Notice that as one increases μ the path of $F_{t,\mu}$ becomes smoother. It actually approaches the normalized unconditional mean of $L_t^{m'} - L_t^m$ as $\mu \to 1$. On the other hand, the local means become very sensitive to specific observations for small values of μ. Figures analogous to Figures 8 and 15 with different values of μ are available upon request.

considered. The coefficient on "After 2008" is positive and statistically significant even after we control for the ZLB period. That is formal evidence that violations of the spanning hypothesis through the medium-run factor became stronger after 2008.

7 Conclusion

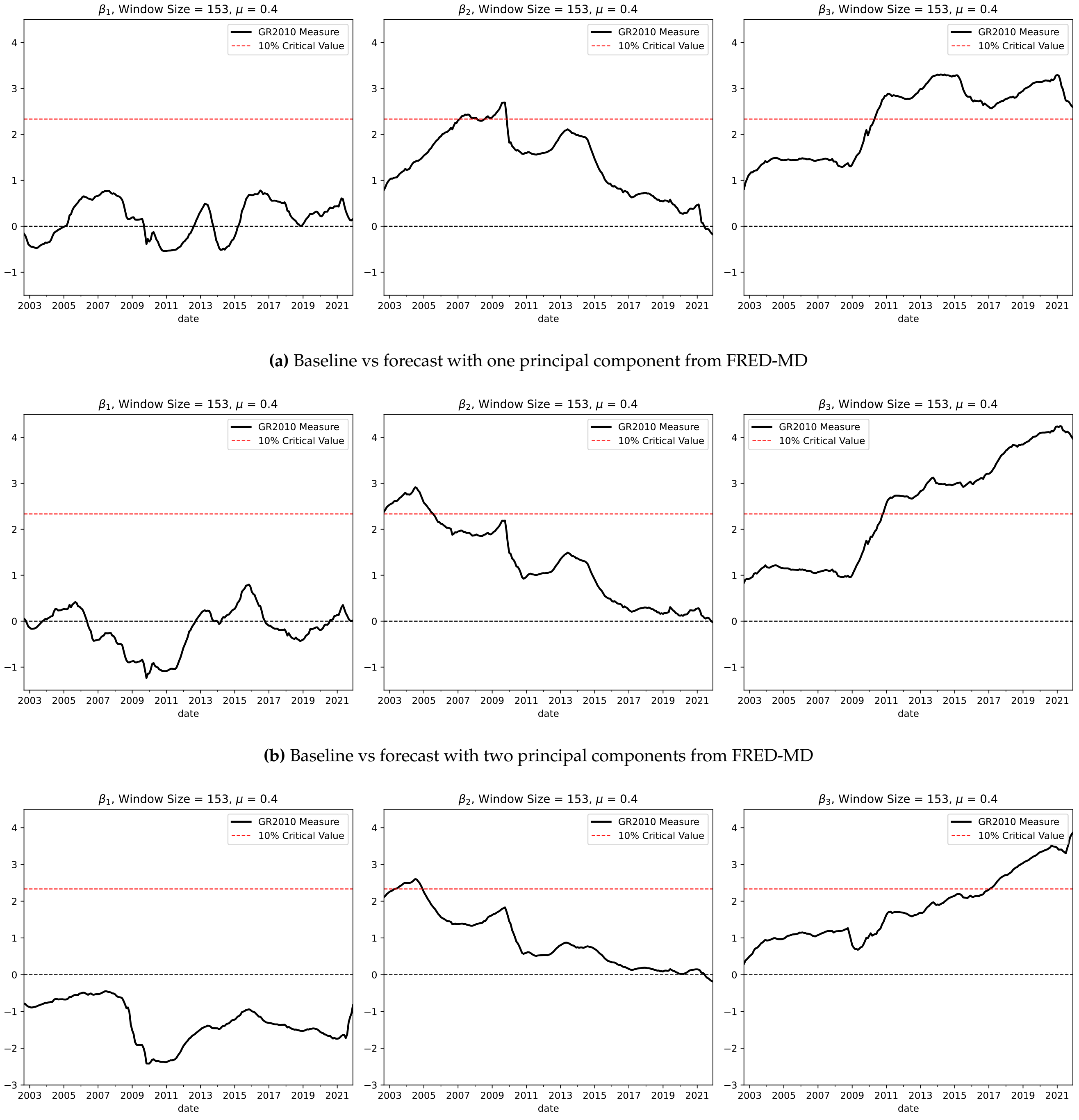
We have studied violations of the spanning hypothesis through the predictability of different parts of the yield curve. Using a Nelson-Siegel model, we showed that violations of the spanning hypothesis are concentrated at the shorter end of the spectrum of maturities, which also leads to an asymmetry in the predictability of excess bond returns. We introduced information from macroeconomic variables to our forecasting designs using both principal components and regularization methods crafted to handle large-dimensional data sets and showed that both methodologies lead to the same conclusion. Across our tests, we found little to no evidence of violations of the spanning hypothesis in the longer end of the maturity spectrum.

We have also shown that these violations at the short end are somewhat predictable. We found evidence that variables related to the business cycle and the current state of monetary policy can predict violations. Both findings are consistent with prior evidence from Borup et al. (2023) and Cieslak (2018). 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.

Even though we deem our results robust across different forecasting designs, our methodology has some caveats. First, all our forecasting designs have strongly relied on linear models. A natural question is whether our results also appear when one allows for non-linearities. Natural approaches would be using random forests and small-scale neural networks since Gu et al. (2020) and Bianchi et al. (2021) show that these models can beat linear regressions in predicting returns of stocks and bond. Another extension would be including the expectations of professional forecasters in the list of macroeconomic variables. These variables are less informative about the business cycle but carry important forward-looking information and they are not included in the FRED-MD dataset.

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 macrofinance 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.

Figure 8: Evolution of $F_{t,\mu}$ for each factor and each forecast model. Each column is dedicated to one of the Nelson-Siegel factors. The baseline forecast is always the one that uses only information from the first principal component of the forward rates. From top to bottom, we add principal components from the FRED-MD data set (one, two, three PCs, respectively). These principal components are taken in real time to avoid any look-ahead bias. The out-of-sample period is January 1990 to December 2021. These rolling means are based on 153 observations ($\mu = 0.4$). The plotting convention is that $F_{t,\mu}$ uses information up to exactly time t.



(a) Baseline vs forecast with one principal component from FRED-MD

(b) Baseline vs forecast with two principal components from FRED-MD

(c) Baseline vs forecast with three principal component from FRED-MD

Table 10: Auxiliary regressions for the conditional predictive ability test from Giacomini and White (2006), as described in (17). The last row displays the *p*-value of the test. We consider here the forecast made with two principal components of the FRED-MD data set. The dependent variable is the difference between the squared prediction error of the baseline model and the model with two principal components from the macro data. Standard errors are computed using Newey and West (1987). The out-of-sample period is January 1990 to December 2021. The first three columns concern the long-run factor, the three middle ones focus on the short-run factor and the last ones are dedicated to the medium-run factor. Stars denote significance at 10%, 5% and 1% respectively. "After 2008" is a dummy variable that assumes for years strictly after 2008 and zero otherwise. "ZLB" is a dummy that is one for the months of Zero Lower Bound and zero otherwise. The baseline forecast is the one that uses only information from the first principal component of the forward rates. The panels condition the competing forecasts on macro data using 1, 2 and 3 principal components of the FRED-MD data set, respectively.

Panel A: Controlling for One PC from Macro Variables

	ΔL_1 (1)	ΔL_1 (2)	ΔL_1 (3)	ΔL_2 (1)	ΔL_2 (2)	ΔL_2 (3)	ΔL_3 (1)	ΔL_3 (2)	ΔL_3 (3)
constant	0.00	0.00	0.00	0.63	0.60	0.63	0.56	1.11**	0.56
	(0.05)	(0.04)	(0.05)	(0.48)	(0.43)	(0.48)	(0.53)	(0.50)	(0.53)
After 2008	0.03		0.00	-0.29		-0.17	2.52***		2.75**
	(0.09)		(0.11)	(0.90)		(1.09)	(0.95)		(1.10)
ZLB		0.04	0.04		-0.35	-0.21		1.81	-0.39
		(0.12)	(0.15)		(0.83)	(0.92)		(1.14)	(1.34)
N	384	384	384	384	384	384	384	384	384
R2 Adj.	0.00	0.00	0.00	0.00	0.00	0.00	0.09	0.03	0.09
GW p-values	0.92	0.90	0.98	0.41	0.41	0.61	0.02	0.02	0.04

Panel B: Controlling for Two PCs from Macro Variables

	ΔL_1 (1)	ΔL_1 (2)	ΔL_1 (3)	ΔL_2 (1)	ΔL_2 (2)	ΔL_2 (3)	ΔL_3 (1)	ΔL_3 (2)	ΔL_3 (3)
constant	0.00	-0.01	0.00	1.70***	1.36***	1.70***	0.88***	1.58***	0.88***
	(0.06)	(0.05)	(0.06)	(0.47)	(0.42)	(0.47)	(0.34)	(0.47)	(0.34)
After 2008	-0.00		-0.06	-1.66*		-1.43	3.30***		2.95**
	(0.11)		(0.12)	(0.92)		(0.97)	(1.05)		(1.49)
ZLB		0.06	0.11		-1.52	-0.44		2.90**	0.66
		(0.16)	(0.18)		(1.00)	(1.02)		(1.33)	(1.86)
N	384	384	384	384	384	384	384	384	384
R2 Adj.	0.00	0.00	0.01	0.05	0.03	0.05	0.16	0.09	0.16
GW p-values	1.00	0.91	0.92	0.02	0.02	0.05	0.00	0.00	0.00

Panel C: Controlling for Three PCs from Macro Variables

	ΔL_1 (1)	ΔL_1 (2)	ΔL_1 (3)	ΔL_2 (1)	ΔL_2 (2)	ΔL_2 (3)	ΔL_3 (1)	ΔL_3 (2)	ΔL_3 (3)
constant	-0.15*	-0.27**	-0.15*	2.21***	1.90***	2.21***	0.65	1.21**	0.65
	(0.09)	(0.12)	(0.09)	(0.67)	(0.54)	(0.67)	(0.54)	(0.58)	(0.54)
After 2008	-0.08		-0.49	-2.33*		-1.31	3.07***		2.35*
	(0.19)		(0.31)	(1.24)		(0.91)	(0.99)		(1.31)
ZLB		0.40*	0.78**		-2.90*	-1.90		3.12**	1.34
		(0.22)	(0.37)		(1.72)	(1.62)		(1.30)	(1.72)
N	384	384	384	384	384	384	384	384	384
R2 Adj.	0.00	0.03	0.07	0.07	0.07	0.09	0.10	0.07	0.11
GW p-values	0.19	0.09	0.19	0.03	0.01	0.03	0.02	0.02	0.04

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.

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(none).

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*.

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. *The 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.

Clark, T. and West, K. (2007). Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics*, 138(1):291–311.

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.

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.

Farmer, L., Schmidt, L., and Timmermann, A. (2022). Pockets of predictability. *SSRN Electronic Journal*.

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. *The Journal of Finance*, 75(3):1327–1370.

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

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*.

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. *The Review of Financial Studies*, 33(5):2223–2273.

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.

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

Litterman, R. B. and Scheinkman, J. (1991). Common factors affecting bond returns. *The 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. *The 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.

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.

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.

A Tables

Table 11: 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 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

Table 12: In-sample predictive regressions targeting the level of the factors as in (8). 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") and the limit case when we include all macroeconomic variables ("All 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	-0.17***	-0.17***	-0.18***	-0.19***	-0.28***	-0.30***	-0.30***	-0.29***	-0.26***	-0.28***	-0.29***	-0.29***
	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
PC 2		0.04	0.05	0.06*		-0.09**	-0.08*	-0.10**		-0.11*	-0.08	-0.09*
		(0.03)	(0.04)	(0.03)		(0.04)	(0.04)	(0.04)		(0.06)	(0.06)	(0.05)
PC 3		-0.01	-0.00	0.01		0.15***	0.15***	0.14***		0.17***	0.18***	0.18***
		(0.03)	(0.03)	(0.03)		(0.04)	(0.05)	(0.04)		(0.06)	(0.05)	(0.05)
PC 4			0.02	0.02			0.03	0.03			0.09*	0.09**
			(0.03)	(0.03)			(0.05)	(0.04)			(0.05)	(0.05)
PC 5			0.11***	0.13***			0.05	0.02			0.17**	0.15**
			(0.04)	(0.04)			(0.05)	(0.05)			(0.07)	(0.07)
PC 6				-0.10***				0.10*				0.03
				(0.04)				(0.05)				(0.09)
PC 7				-0.02				-0.12***				-0.14**
				(0.03)				(0.04)				(0.05)
PC 8				-0.03				-0.19***				-0.23***
				(0.04)				(0.07)				(0.08)
R-squared Adj.	0.86	0.86	0.87	0.87	0.35	0.40	0.40	0.44	0.41	0.45	0.47	0.49
N	588	588	588	588	588	588	588	588	588	588	588	588
R-squared Adj. (No Macro)	0.82	0.82	0.82	0.82	0.07	0.07	0.07	0.07	0.29	0.29	0.29	0.29
R-squared Adj. (All Macro)	0.91	0.91	0.91	0.91	0.64	0.64	0.64	0.64	0.64	0.64	0.64	0.64

Table 13: OOS R^2 using the random walk as benchmark. 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 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 linear models in (2) to make the forecasts. Principal components are extracted in a sequential way so they introduce no look-ahead bias. Panel A lets C_t store the three lagged Nelson-Siegel factors while Panel B controls for the forward rates. 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: Controlling for Lagged Betas

	R^2_{OOS}						p-values			
Target	No Macro	1 PC	5 PC	8 PC	10 PC	All Macro	1 PC	5 PC	8 PC	10 PC
Beta 1	-0.77	-0.25	-0.30	-0.25	-0.36	-2.82	0.01	0.03	0.01	0.04
Beta 2	-0.55	-0.36	0.05	0.06	0.06	-0.78	0.05	0.00	0.01	0.01
Beta 3	-0.59	-0.48	-0.36	-0.36	-0.37	-1.04	0.14	0.05	0.06	0.08
Innovation 1	-0.22	-0.11	-0.05	-0.06	-0.08	-1.58	0.06	0.02	0.04	0.13
Innovation 2	0.13	0.01	0.19	0.21	0.20	-0.77	1.00	0.22	0.20	0.17
Innovation 3	0.01	0.01	-0.03	-0.04	-0.07	-1.47	0.51	0.72	0.77	0.87

Panel B: Controlling for Forward Rates

	R^2_{OOS}						p-values			
Target	No Macro	1 PC	5 PC	8 PC	10 PC	All Macro	1 PC	5 PC	8 PC	10 PC
Beta 1	-0.18	-0.18	-0.16	-0.10	-0.09	-1.98	0.57	0.36	0.13	0.18
Beta 2	-0.01	0.03	0.24	0.23	0.25	-0.66	0.18	0.03	0.06	0.05
Beta 3	-0.33	-0.35	-0.21	-0.22	-0.25	-1.20	0.68	0.11	0.17	0.25
Innovation 1	-0.12	-0.09	-0.07	0.02	0.04	-1.71	0.16	0.24	0.05	0.09
Innovation 2	0.08	0.10	0.20	0.17	0.19	-0.66	0.26	0.19	0.28	0.24
Innovation 3	-0.03	-0.04	0.02	0.02	0.00	-1.34	0.67	0.19	0.25	0.36

Table 14: R^2_{OOS} from different models and for different targets, using the random walk the benchmark. 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 lets C_t store the three lagged Nelson-Siegel factors while Panel B controls for the forward rates. 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 Lagged Betas

Target	No Macro Data			All Macro Data			p-value		
	Ridge	Lasso	Elastic Net	Ridge	Lasso	Elastic Net	Ridge	Lasso	Elastic Net
Beta 1	-5.47	-5.51	-5.90	-4.48	-4.33	-4.76	0.03	0.00	0.00
Beta 2	-0.11	-0.08	-0.09	0.12	0.12	0.10	0.04	0.03	0.05
Beta 3	-0.84	-0.86	-0.88	-0.63	-0.49	-0.52	0.14	0.01	0.01
Innovation 1	-0.05	-0.00	-0.00	-0.38	-0.02	-0.00	0.99	0.59	0.48
Innovation 2	0.14	0.10	0.12	0.22	0.23	0.22	0.13	0.00	0.02
Innovation 3	0.03	0.01	0.02	-0.19	0.04	0.05	0.95	0.35	0.24

Panel B: Conditioning on Lagged Forward Rates

Target	No Macro Data			All Macro Data			p-value		
	Ridge	Lasso	Elastic Net	Ridge	Lasso	Elastic Net	Ridge	Lasso	Elastic Net
Beta 1	-4.88	-5.05	-5.26	-4.27	-4.43	-4.46	0.03	0.01	0.00
Beta 2	-0.02	-0.04	-0.05	0.10	0.15	0.13	0.14	0.02	0.03
Beta 3	-0.63	-0.69	-0.70	-0.44	-0.44	-0.43	0.04	0.01	0.01
Innovation 1	0.10	0.11	0.08	0.02	0.13	0.14	0.89	0.37	0.09
Innovation 2	-0.02	-0.04	-0.06	0.12	0.20	0.17	0.03	0.00	0.00
Innovation 3	0.04	0.02	-0.03	-0.11	-0.05	-0.04	0.99	0.91	0.68

Table 15: Auxiliary regressions for the conditional predictive ability test from Giacomini and White (2006), as described in (17). The last row displays the p-value of the test. We consider here the forecast made with one principal components of the FRED-MD data set. The dependent variable is the difference between the squared prediction error of the baseline model and the model with one principal component from the macro data. Standard errors are computed using Nelson and Siegel (1987). The out-of-sample period is January 1990 to December 2021. The first three columns concern the long-run factor, the three middle ones focus on the short-run factor and the last ones are dedicated to the medium-run factor. Stars denote significance at 10%, 5% and 1% respectively.

	ΔL_1 (1)	ΔL_1 (2)	ΔL_1 (3)	ΔL_2 (1)	ΔL_2 (2)	ΔL_2 (3)	ΔL_3 (1)	ΔL_3 (2)	ΔL_3 (3)
constant	0.15	0.36**	0.27	0.13	0.14	0.59	0.28	-0.26	0.86*
	(0.43)	(0.16)	(0.31)	(0.45)	(0.16)	(0.38)	(0.45)	(0.28)	(0.44)
NBER	-1.13***		-0.54**	-0.32		-0.39	1.25***		0.46
	(0.28)		(0.24)	(0.55)		(0.40)	(0.46)		(0.36)
IP	-0.21**		-0.18**	0.29***		0.29***	0.19*		0.15
	(0.10)		(0.09)	(0.07)		(0.08)	(0.10)		(0.09)
Unemployment	-0.03		-0.03	-0.09*		-0.21***	-0.03		-0.25**
	(0.07)		(0.06)	(0.06)		(0.06)	(0.09)		(0.11)
Inflation	0.08		0.15	0.17		0.37***	-0.11		0.10
	(0.06)		(0.10)	(0.11)		(0.11)	(0.08)		(0.12)
Fed Funds		-0.03	-0.07*		-0.01	-0.13***		-0.05	-0.10**
		(0.03)	(0.04)		(0.04)	(0.03)		(0.06)	(0.05)
MP Cycle		0.20***	0.16**		0.14	0.02		-0.10**	-0.14***
		(0.06)	(0.07)		(0.09)	(0.07)		(0.05)	(0.05)
MP Activity		-0.51*	-0.45*		-0.09	0.18		0.79**	0.86***
		(0.26)	(0.27)		(0.41)	(0.27)		(0.36)	(0.23)
ZLB		-0.17	-0.03		-0.17	0.49**		0.51	1.12**
		(0.31)	(0.32)		(0.23)	(0.24)		(0.37)	(0.53)
N	384	384	384	384	384	384	384	384	384
R2 Adj.	0.11	0.15	0.19	0.22	0.05	0.31	0.12	0.16	0.29
GW p-value	0.08	0.09	0.20	0.05	0.19	0.12	0.04	0.06	0.13

Table 16: Auxiliary regressions for the conditional predictive ability test from Giacomini and White (2006), as described in (17). The last row displays the *p*-value of the test. We consider here the forecast made with three principal components of the FRED-MD data set. The dependent variable is the difference between the squared prediction error of the baseline model and the model with three principal components from the macro data. Standard errors are computed using Nelson and Siegel (1987). The out-of-sample period is January 1990 to December 2021. The first three columns concern the long-run factor, the three middle ones focus on the short-run factor and the last ones are dedicated to the medium-run factor. Stars denote significance at 10%, 5% and 1% respectively.

	ΔL_1 (1)	ΔL_1 (2)	ΔL_1 (3)	ΔL_2 (1)	ΔL_2 (2)	ΔL_2 (3)	ΔL_3 (1)	ΔL_3 (2)	ΔL_3 (3)
constant	-0.69**	0.03	-0.71*	-0.76*	0.16	-0.11	0.65	0.21	1.11***
	(0.34)	(0.16)	(0.38)	(0.44)	(0.24)	(0.35)	(0.43)	(0.24)	(0.39)
NBER	-1.44**		-1.19*	-0.70		-1.57***	-0.05		-0.68
	(0.62)		(0.68)	(0.58)		(0.52)	(0.40)		(0.50)
IP	-0.11		-0.09	0.09		0.04	-0.06		-0.09*
	(0.07)		(0.07)	(0.06)		(0.07)	(0.05)		(0.05)
Unemployment	0.11**		0.10	-0.02		-0.14***	0.01		-0.15**
	(0.06)		(0.06)	(0.06)		(0.05)	(0.07)		(0.06)
Inflation	0.07		0.09	0.38**		0.59***	-0.28***		-0.11
	(0.05)		(0.07)	(0.15)		(0.12)	(0.06)		(0.07)
Fed Funds		-0.01	-0.01		0.03	-0.13***		-0.12***	-0.09*
		(0.03)	(0.04)		(0.05)	(0.05)		(0.04)	(0.05)
MP Cycle		0.17**	0.11**		-0.11	-0.30***		-0.04	-0.13***
		(0.08)	(0.04)		(0.11)	(0.08)		(0.06)	(0.04)
MP Activity		-0.13	0.07		-0.34	0.05		0.27	0.58*
		(0.21)	(0.31)		(0.57)	(0.27)		(0.25)	(0.32)
ZLB		0.36	0.22		-0.67	-0.04		0.28	0.69**
		(0.25)	(0.27)		(0.49)	(0.28)		(0.32)	(0.30)
N	384	384	384	384	384	384	384	384	384
R2 Adj.	0.18	0.10	0.21	0.29	0.09	0.47	0.14	0.15	0.23
GW p-value	0.12	0.37	0.12	0.01	0.02	0.05	0.04	0.07	0.16

B Figures

Figure 9: 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.

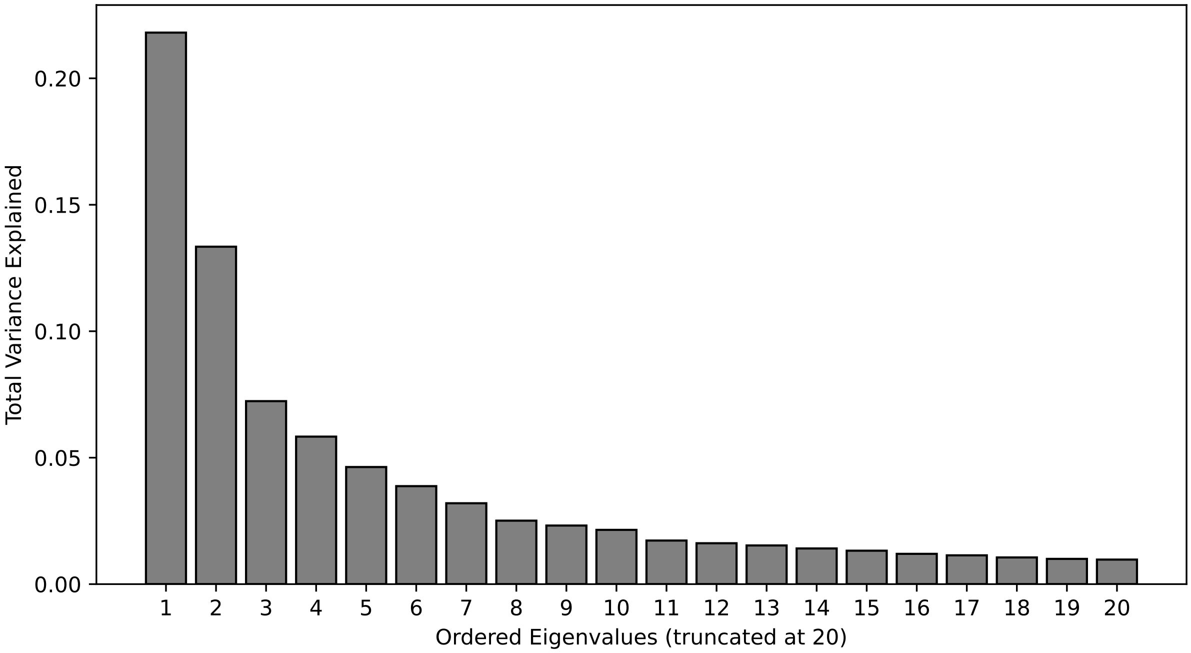


Figure 10: Relative MSE predicting returns using three principal components of the yield curve as controls. For each maturity we show the ratio between the MSE attained with different numbers of principal components from the macroeconomic data and the baseline model that uses information only from the 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.

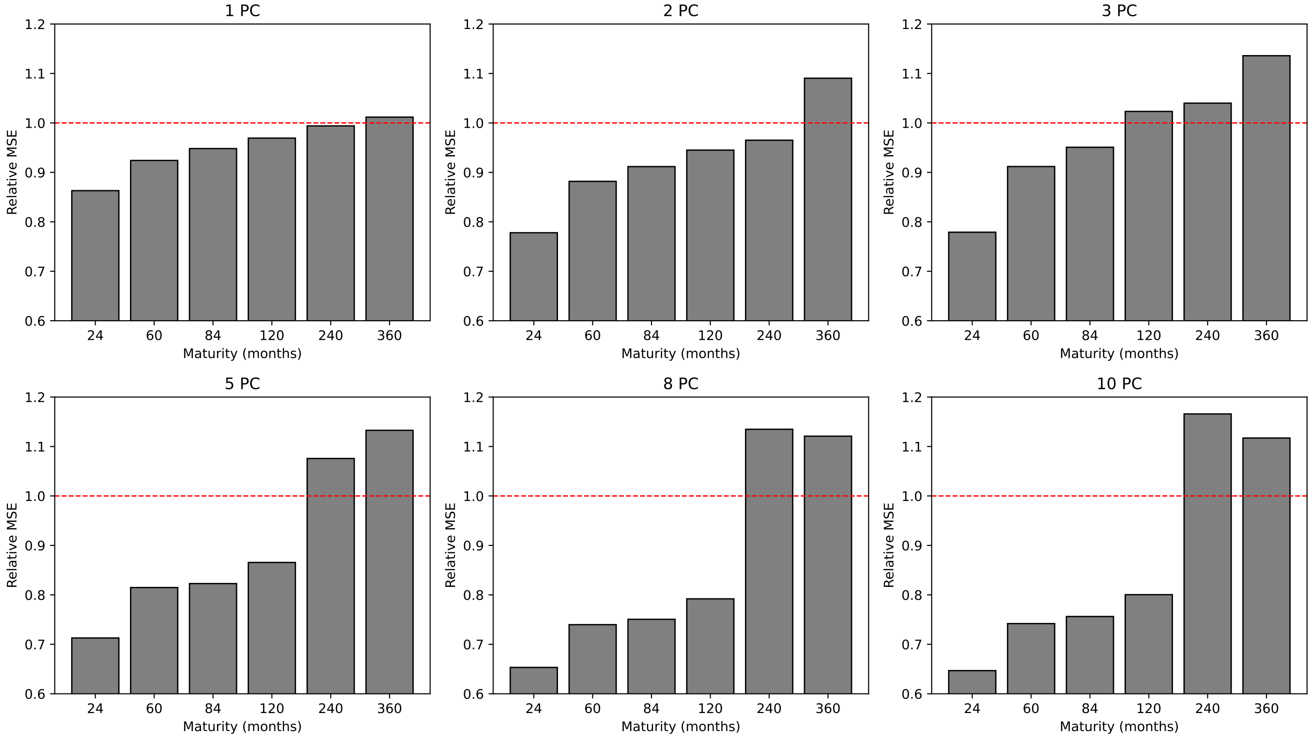


Figure 11: Profiling of the decay parameter. For each value of λ, we fit the Nelson-Siegel model by OLS date by date. Then we compute a monthly measure of the average squared fitting error in the cross-section. We finally average over time and plot this information denoted by "Fit Error" as a function 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.

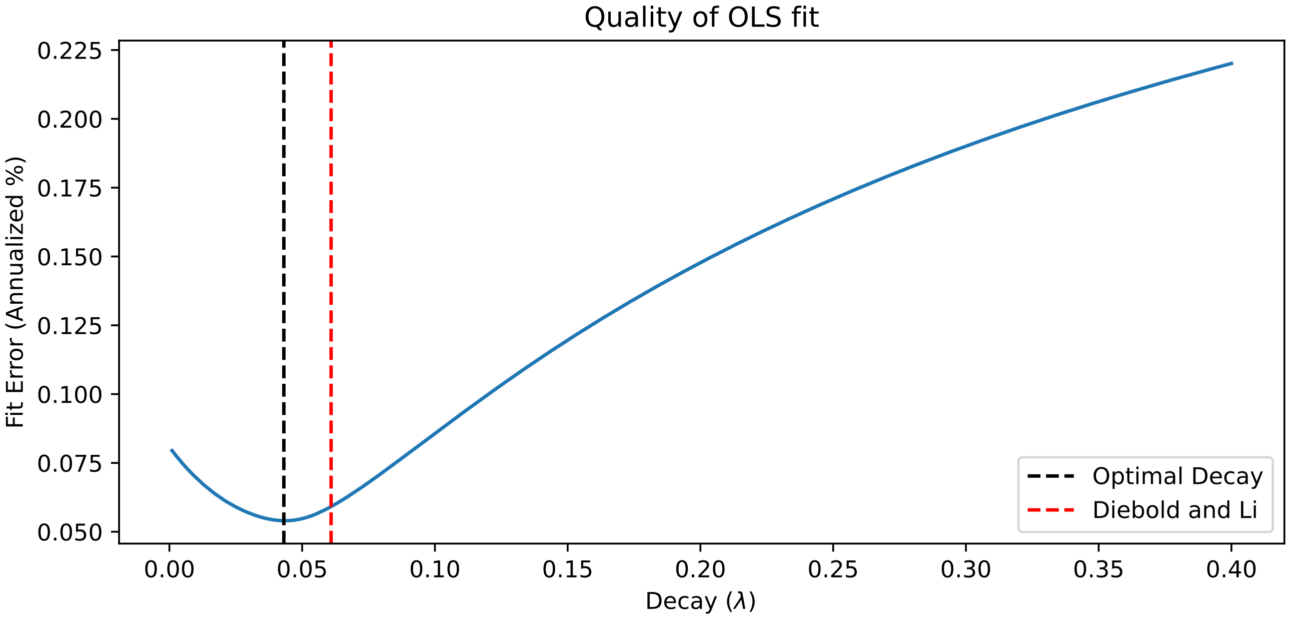


Figure 12: Spectral decomposition of forward rates. The sample ranges from January 1973 to December 2021. We normalize the panel of forward rates and then compute the spectral decomposition of the associated correlation matrix. We display the cumulative variance explained by the eigenvalues.

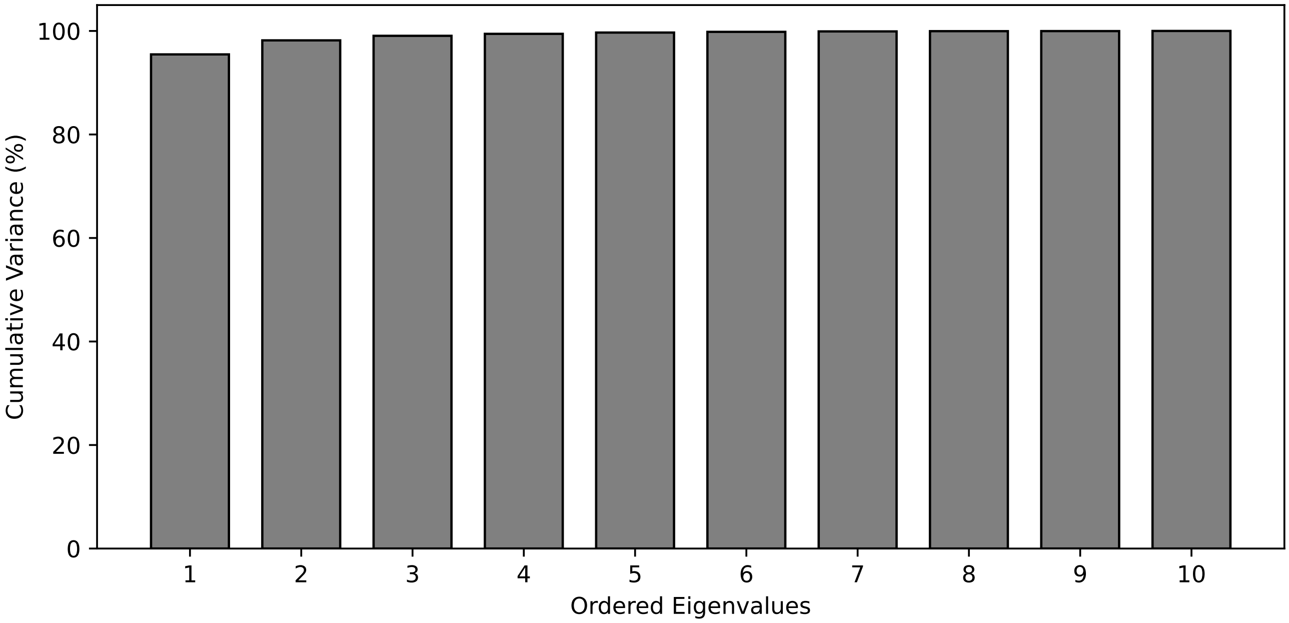


Figure 13: Forward rates and the first principal component. The sample period ranges from January 1973 to December 2021.

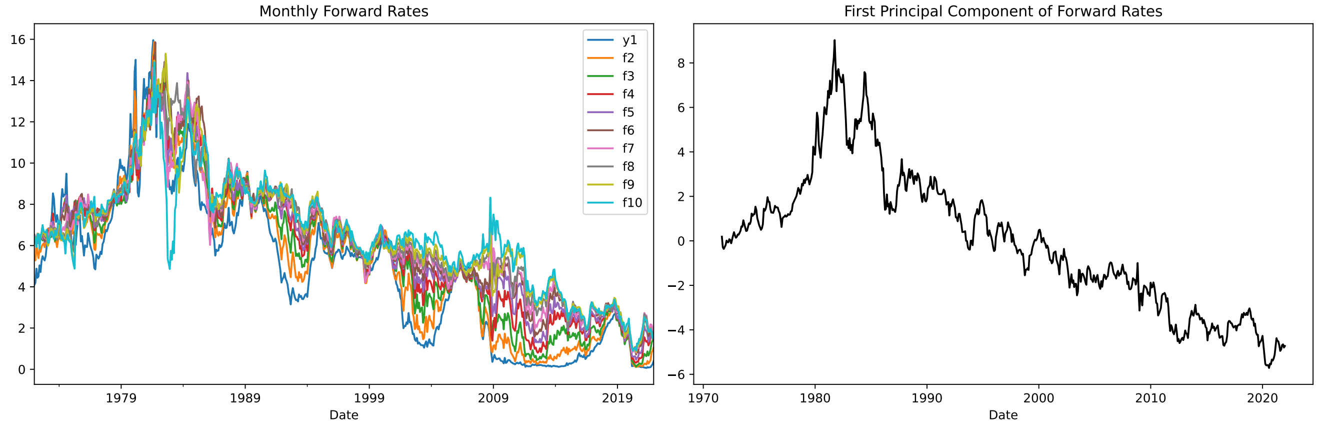


Figure 14: Conditioning variables for the conditional predictive ability test. Refer to the text for the definitions of each variable. The sample ranges from January 1990 to December 2021. All data comes from the St Louis Fed.

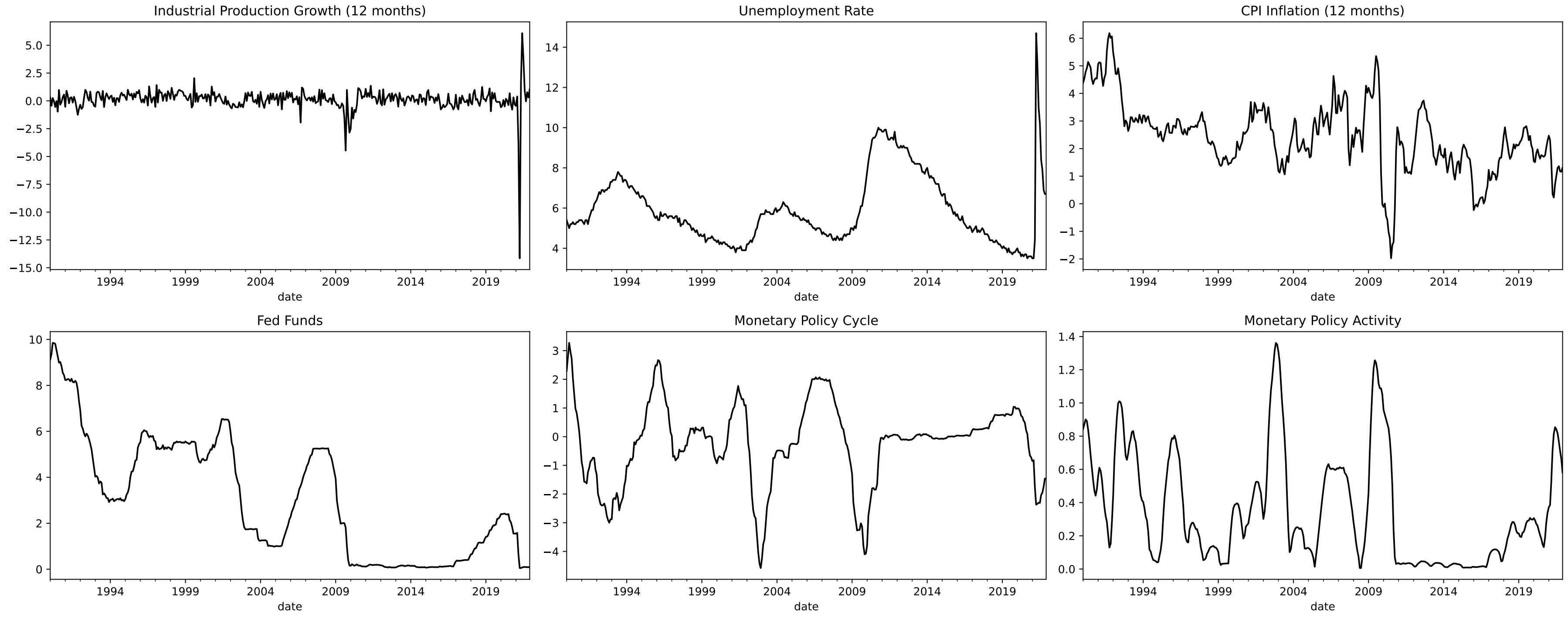
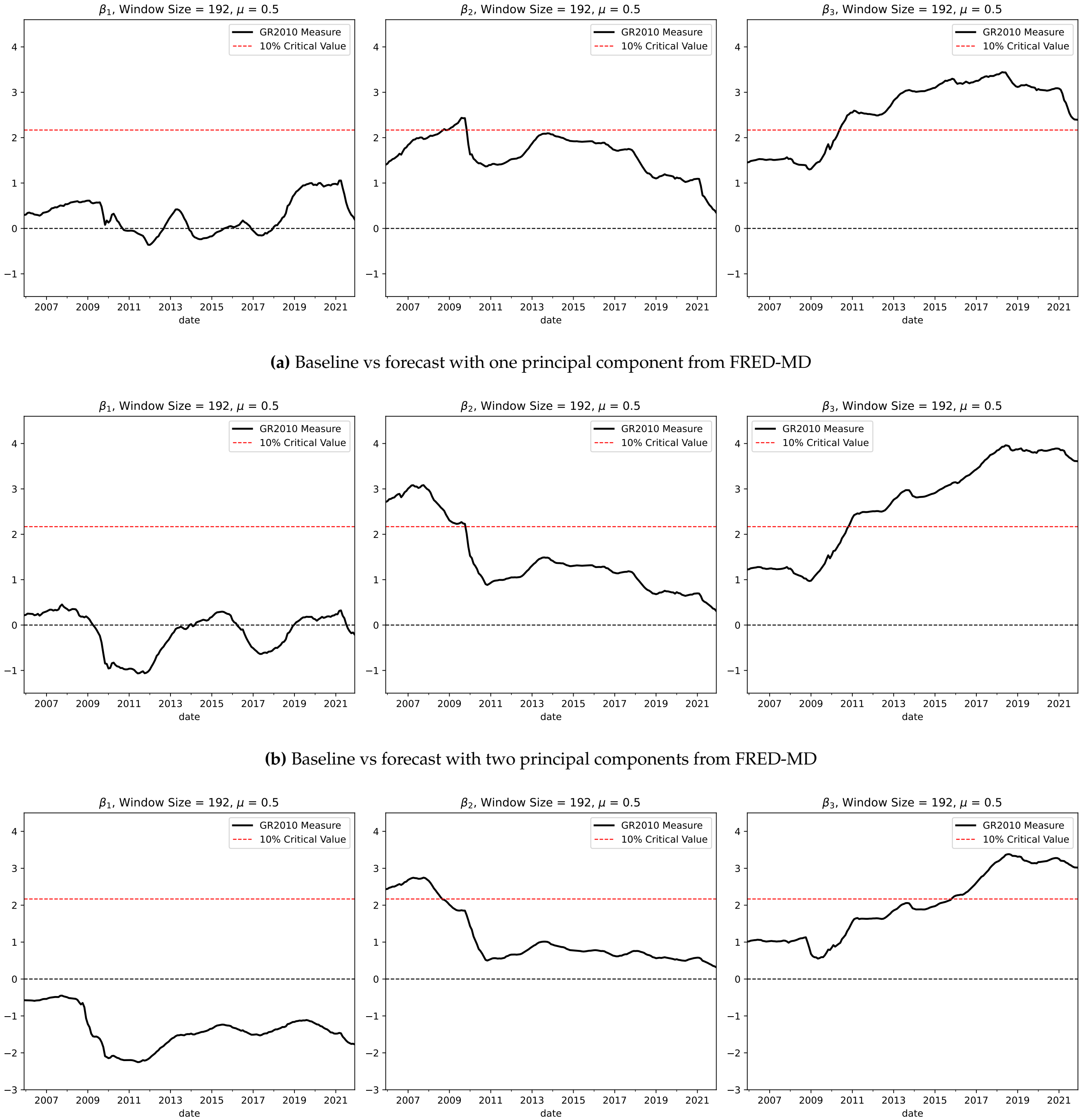


Figure 15: Evolution of $F_{t,\mu}$ for each factor and each forecast model. Each column is dedicated to one of the Nelson-Siegel factors. The baseline forecast is always the one that uses only information from the first principal component of the forward rates. From top to bottom, we add principal components from the FRED-MD data set (one, two, three PCs, respectively). These principal components are taken in real time to avoid any look-ahead bias. The out-of-sample period is January 1990 to December 2021. These rolling means are based on 153 observations ($\mu = 0.5$). The plotting convention is that $F_{t,\mu}$ uses information up to exactly time t.



(a) Baseline vs forecast with one principal component from FRED-MD

(b) Baseline vs forecast with two principal components from FRED-MD

(c) Baseline vs forecast with three principal components from FRED-MD

C FRED-MD Variables

In this appendix we report the full set of variables from the FRED-MD data set we use. Table 17 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 17: 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

D Proof of Proposition 1

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

$$y_t^{(\tau)} = \beta_{1,t} + \beta_{2,t}\left(\frac{1-e^{-\lambda\tau}}{\lambda\tau}\right) + \beta_{3,t}\left(\frac{1-e^{-\lambda\tau}}{\lambda\tau} - e^{-\lambda\tau}\right) \tag{21}$$

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

$$n \cdot \xi_t(n) = n\beta_{1,t} + \left(\frac{1-e^{-\theta n}}{\theta}\right)[\beta_{2,t} + \beta_{3,t}] - n\beta_{3,t}e^{-\theta n} \tag{22}$$

$$(n-1)\cdot \xi_{t+12}(n-1) = (n-1)\beta_{1,t+12} + \left(\frac{1-e^{-\theta(n-1)}}{\theta}\right)[\beta_{2,t+12} + \beta_{3,t+12}] - (n-1)\beta_{3,t+12}e^{-\theta(n-1)} \tag{23}$$

$$\xi_t(1) = \beta_{1,t} + \left(\frac{1-e^{-\theta}}{\theta}\right)[\beta_{2,t} + \beta_{3,t}] - \beta_{3,t}e^{-\theta} \tag{24}$$

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

$$n\beta_{1,t} - (n-1)\beta_{1,t+12} - \beta_{1,t} = (n-1)\left[\beta_{1,t} - \beta_{1,t+12}\right] \tag{25}$$

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

$$\begin{aligned}\left(\frac{1-e^{-\theta n}}{\theta} - \frac{1-e^{-\theta}}{\theta}\right)[\beta_{2,t} + \beta_{3,t}] - \left(\frac{1-e^{-\theta(n-1)}}{\theta}\right)[\beta_{2,t+12} + \beta_{3,t+12}] = \\ \left(\frac{1-e^{-\theta(n-1)}}{\theta}\right)\left[\left(e^{-\theta}\beta_{2,t} - \beta_{2,t+12}\right) + \left(e^{-\theta}\beta_{3,t} - \beta_{3,t+12}\right)\right]\end{aligned} \tag{26}$$

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.