STAFF
REPORTS

NO. 884
APRIL 2019

REVISED
AUGUST 2023

Deconstructing the Yield Curve

Richard K. Crump | Nikolay Gospodinov

FEDERAL RESERVE BANK *of* NEW YORK

Deconstructing the Yield Curve
Richard K. Crump and Nikolay Gospodinov
Federal Reserve Bank of New York Staff Reports, no. 884
April 2019; revised August 2023
JEL classification: G10, G12, C15, C58

Abstract

We introduce a novel nonparametric bootstrap for the nominal yield curve which is agnostic to the true factor structure. We deconstruct the yield curve into primitive objects, with weak cross-sectional and time-series dependence, which serve as building blocks for resampling the data. We analyze the asymptotic and finite-sample properties of the bootstrap for mimicking salient features of the data and conducting inference on bond return predictability. We demonstrate the applicability of our results to: the "tent shape" in forward rates, regression tests of the expectations hypothesis, the role of trend inflation in expected bond returns, and yield-based forecasts of recessions.

Key words: term structure of interest rates, resampling-based inference, factor models, bond risk premiums, predictive regression of bond returns

---

Crump: Federal Reserve Bank of New York (email: richard.crump@ny.frb.org). Gospodinov: Federal Reserve Bank of Atlanta (email: nikolay.gospodinov@atl.frb.org). The authors thank the Editor (Ralph Koijen), an Associate Editor, two anonymous referees, Tobias Adrian, Anna Cieslak (discussant), Mark Fisher, Ken Garbade, Domenico Giannone, Carsten Jentsch (discussant), David Lucca, Ulrich Müller, Stijn Van Nieuwerburgh, Peter Van Tassel, Paula Tkac, Erik Vogt, Desi Volker, Mark Watson, Jonathan Wright, and seminar participants at the 9th Bundesbank Term Structure Workshop, Columbia Business School, the Board of Governors of the Federal Reserve, the Federal Reserve Bank of Atlanta, the Federal Reserve Bank of New York, the 2019 LBS Summer Finance Symposium, McGill University, the 2019 NBER Summer Institute, and Princeton University for helpful discussions and comments. They also thank Oliver Kim for excellent research assistance.

This paper presents preliminary findings and is being distributed to economists and other interested readers solely to stimulate discussion and elicit comments. The views expressed in this paper are those of the author(s) and do not necessarily reflect the position of the Federal Reserve Bank of New York, the Federal Reserve Bank of Atlanta, or the Federal Reserve System. Any errors or omissions are the responsibility of the author(s).

To view the authors' disclosure statements, visit
https://www.newyorkfed.org/research/staff_reports/sr884.html.

1 Introduction

The term structure of interest rates is a key object of study in both macroeconomics and finance. Testing theories of interest-rate determination have important implications for understanding how asset prices reflect expectations about the future, risk preferences, market structure, or other considerations. Term structure modeling is generally parametric in nature and reduced-form inference tends to rely on off-the-shelf procedures designed for generic regression specifications. However, accurate inference is a formidable task as this parametric setting presents a confluence of econometric challenges.

First, parametric models of the term structure of interest rates require the correct specification of the underlying factor structure of the data. Recent research highlights the challenge of identifying the correct factor space. Uhlig (2009) and Onatski and Wang (2021) demonstrate that high time-series persistence can induce spurious commonality across series suggestive of a stronger factor structure than might exist. Crump and Gospodinov (2021) show that, even in the absence of strong serial correlation, certain properties of maturity-ordered assets induce standard metrics to possibly favor a much lower dimension than the true dimension of the underlying factor space. Other work (e.g., Duffee (2011), Joslin et al. (2014)) has argued for the presence of "hidden" or unspanned factors which represent information that is not fully incorporated in bond yields.

Second, even under correct specification, the underlying yield factors will exhibit a high degree of time-series persistence which presents a perennial obstacle to trustworthy inference in time-series analysis. A number of papers, starting with Cavanagh et al. (1995), show that estimation and inference of return predictability is sensitive to the exact degree of persistence and can result in severe bias and size distortions (see also, e.g., Stambaugh (1999), Ferson et al. (2003), Jansson and Moreira (2006), Campbell and Yogo (2006), Wei and Wright (2013), Bauer and Hamilton (2018)).

Third, unlike most empirical settings where there are distinct dependent and explanatory variables, inference in fixed-income regressions is often conducted on regression coefficients where both regressors and the regressand are linear combinations of the same underlying yield curve. This makes the setting distinct from the usual linear regression setup and implies that conventional resampling approaches will produce bootstrapped samples which do not satisfy economic relations and identities. Thus, the conventional parametric framework can omit potentially important factors and key aspects of their dynamics, and generate yields with internally inconsistent properties.

In this paper, we introduce a novel nonparametric resampling procedure which is tailored to assets with a finite maturity structure such as the nominal yield curve. Rather than resampling yields or prices directly, we instead resample excess returns along with a single, far-in-the-future forward rate. This can be thought of as an economically-motivated transformation of the original price data that produces primitive objects with more appealing statistical properties which, in turn, serve as the basis for bootstrapping the data. We stress that our proposed method differs conceptually and practically from the typical approach (e.g., Giglio and Kelly (2017), Bauer and Hamilton (2018)) that specifies yields as the primitive process, takes a stand on their factor structure and its law of motion, and employs a fully parametric model of yields in terms of their assumed pricing factors. In contrast, our method is nonparametric which enables valid resampling of yield curves while remaining agnostic about the underlying data generating process. Our bootstrap procedure does not require as an input either the correct factor structure or the true pricing model which generated the data; moreover, it is robust to unaccounted forms of time-series and cross-sectional dependence and presence of conditional heteroskedasticity. We prove that our bootstrap procedure is asymptotically valid under mild assumptions. Importantly, our analyses and results translate directly to all other assets with a finite maturity structure such as options, swaps, or futures.

We assess the properties of our bootstrap in replicating some salient features of the data (such the "tent shape" of Cochrane and Piazzesi (2005)) as well as conducting inference in the context of bond return predictability regressions. A key feature of our bootstrap is that the resampled data naturally satisfies term structure identities ensuring that any predictability in future returns from past yields or forwards is retained. The Monte Carlo experiments illustrate the ability of our method to retain the unknown factor structure in the data and show that the resulting bootstrap-based inference controls size well for multi-period holding returns at various maturities. The simulation evidence suggests that our bootstrap is well-suited for settings where the predictors of interest are functions of the yield curve themselves (e.g., Fama and Bliss (1987), Campbell and Shiller (1991), Cochrane and Piazzesi (2005, 2008)) or are augmented with other external (macroeconomic) predictors (e.g., Cooper and Priestley (2008), Ludvigson and Ng (2009), Joslin et al. (2014), Cieslak and Povala (2015), Ghysels et al. (2018), Haddad and Sraer (2020), Bauer and Rudebusch (2020)).

In the empirical analysis, we consider three main applications of our new bootstrap method.

First, we revisit the regression-based tests of the expectations hypothesis. We show how to view all of the existing specifications for testing the expectations hypothesis in terms of primitive objects (difference returns) and introduce a novel regression formulation which is better suited to the time-series properties of bond prices. Since our bootstrap is not model specific, it offers a unified framework for evaluating the empirical relevance of the expectations hypothesis across different specifications. Overall, we find some evidence against the expectations hypothesis but it is generally weak.

The second application provides stronger evidence against the expectations hypothesis by investigating the predictability of bond risk premia using trend inflation as suggested by Cieslak and Povala (2015). We show how our nonparametric bootstrap can accommodate the highly persistent nature of trend inflation by preserving its potential relationship with bond data. Moreover, the sampling uncertainty from the construction of trend inflation as a generated regressor is properly reflected in our inference procedure. Our results support the conclusions of Cieslak and Povala (2015) regarding the importance of trend inflation in driving the low-frequency movements in bond yields, by further extending the sample until the end of 2022.

Finally, we use our new bootstrap procedure to construct bias-corrected estimates and confidence intervals for the probability of recession driven by the slope of the yield curve. We show that for small fitted probabilities – periods associated with a wider term spread – the bias-corrected estimate is below the standard estimate; in contrast, for large fitted probabilities – periods associated with a compressed term spread – the bias-corrected estimate is comfortably above it. The direction of the bias correction closely aligns with the impressive forecasting record associated with the term spread over the last 50 or so years and highlights the benefits of our bootstrap approach. We find that, as of the first quarter of 2023, the probability of a contraction in real GDP growth in 2023-2024 is elevated – even after taking the sampling uncertainty into account – and is above the peaks in the last four cycles.

This paper is organized as follows. In Section 2 we summarize key properties of maturity-ordered assets which highlight the pitfalls in committing to a tightly parameterized, finite-dimensional factor structure to conduct inference. This section also provides additional motivation for a nonparametric bootstrap procedure based on results in Cochrane and Piazzesi (2005) and associated simulation experiments. In Section 3 we introduce the nonparametric bootstrap method; we establish its asymptotic validity under general assumptions and demonstrate its appealing finite-sample properties for inference in bond return predictability regressions with external regressors. Section 4 presents results from our empirical applications. In Section 4.1, we revisit regression-based tests of the expectations hypothesis and use our bootstrap to uncover key statistical features of this approach. We show how to tailor our bootstrap to applications with generated regressors in Section 4.2 and investigate the predictive power of trend inflation as in Cieslak and Povala (2015). In Section 4.3 we construct bootstrap-based, bias-corrected estimates and confidence intervals for the probability of contractions in real GDP growth based on the term spread and other aspects of the term structure of interest rates. Section 5 concludes. Proofs of the main results and data description are collected in Appendices A and B, respectively. Additional results are presented in a Supplemental Appendix (hereafter, "SA").

2 The Need for Robust Inference: A Motivation

As argued in the introduction, when analyzing the term structure of interest rates, it is desirable to resort to inference methods that exhibit robustness to uncertainty in the factor structure and time-series properties of the bond data. We deal with both of these problems by deconstructing the yield curve into primitive objects, characterized by weak cross-sectional and time-series dependence, that represent the returns on a forward trade. Resampling of these primitive objects and then reconstructing the forward and yield curves via definitional identities allows us to preserve the main features of the original data. We illustrate the advantages of this approach in different setups and contrast its properties to a parametric (model-based) bootstrap as well as standard asymptotic methods for inference.

2.1 Uncertainty about Cross-Sectional Dependence

We start by highlighting some challenges in characterizing the underlying factor structure of bond data. Define $p_t^{(n)}$ as the time t log price of a zero-coupon bond with n periods to maturity which pays \$1 at time $t+n$, where $t = 1, \ldots, T$ and $n = 1, ..., N$. The corresponding log yield is denoted by $y_t^{(n)}$ and satisfies $p_t^{(n)} = -ny_t^{(n)}$. The log forward rate corresponding to a one-period investment between $t+n-1$ and $t+n$, $f_t^{(n)}$, is defined as

$$f_t^{(n)} \equiv p_t^{(n-1)} - p_t^{(n)}. \tag{1}$$

Since $p_t^{(1)} = 0$, the one-period rate, $y_t^{(1)}$, may equivalently be written as $f_t^{(1)}$. Using the recursive (in n) nature of equation (1) and the definition of yields, we have

$$p_t^{(n)} = -\sum\nolimits_{i=1}^{n} f_t^{(i)}, \qquad y_t^{(n)} = \frac{1}{n}\sum\nolimits_{i=1}^{n} f_t^{(i)}. \tag{2}$$

We observe that $p_t^{(n)}$ and $y_t^{(n)}$ are cross-sectional partial sums and partial averages of forward rates, respectively. These formulas demonstrate that two yields, $y_t^{(n)}$ and $y_t^{(m)}$, have $\min(m,n)$ forwards in common (and the same for prices). This overlap, which arises solely from these term-structure identities, implies differential behavior in the covariance or correlation matrix of forwards relative to yields (or prices); in fact, prices, will always show stronger local correlation than forwards across the same maturities.[1]

What are the implications of these cross-sectional restrictions for characterizing the true factor space? To answer this question, it is instructive to contrast these definitional relationships with the current practice of extracting principal components (PCs) from yields and some stylized facts about the estimated factor loadings. It is now ubiquitous to estimate the factor structure of the term structure by applying static principal components analysis to *yields* across maturities. The top left plot of Figure 1 presents the factor loadings for yields that are widely documented in the term structure literature.[2] In this particular sample, these first 3 principal components explain almost 100% of the cross-sectional variation of yields with only the first principal component explaining in excess of 98.6% of the variation. At first glance, this is suggestive of a low-dimensional factor structure.

Moreover, a similar pattern appears to hold for any asset with a finite maturity structure. In the top right and bottom left panel we show the associated PC loadings for oil futures and S&P 500 options returns.[3] However, in the last panel we show the PC loadings from a panel data set of the seasonal cycle of global surface temperature. Remarkably again, the loadings are very similar in

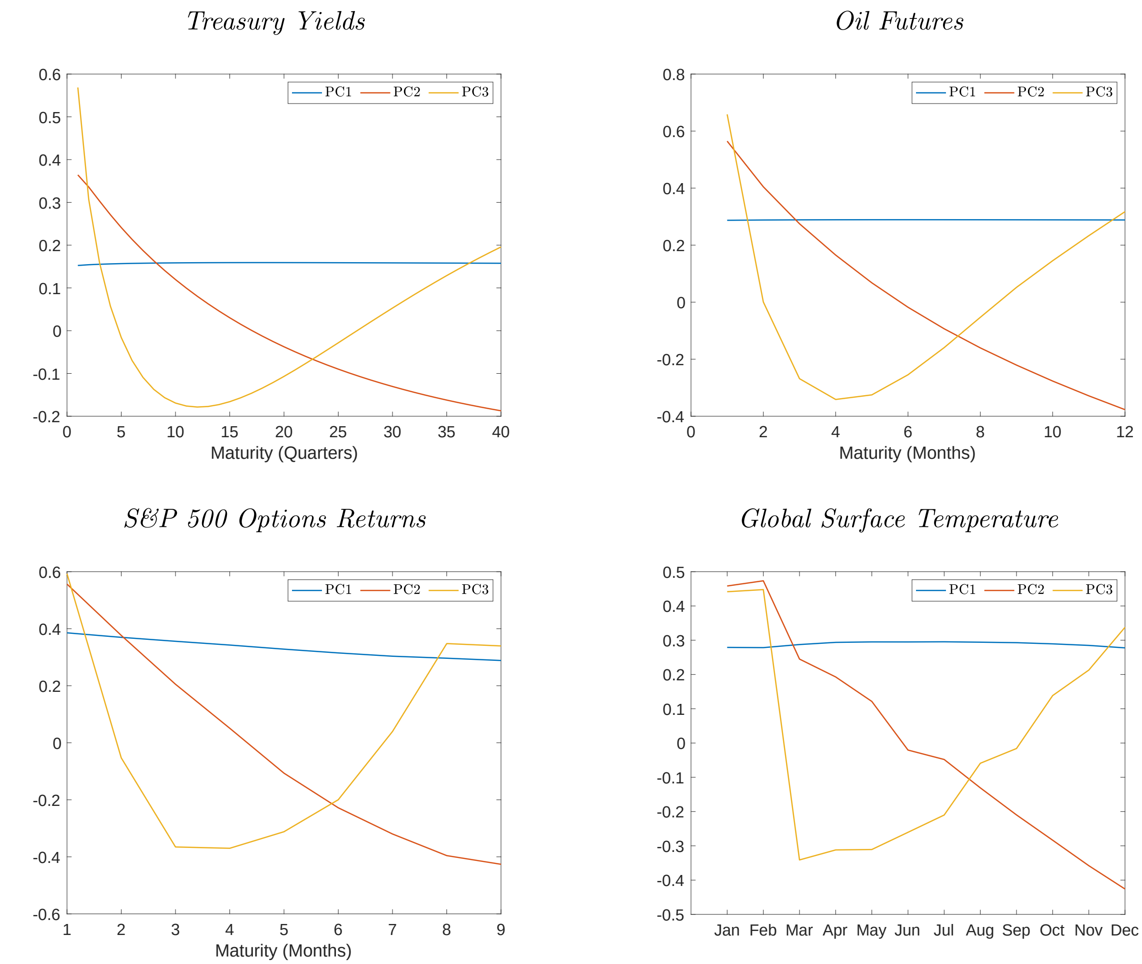
[1] This is true provided forwards are positively correlated across maturities (as is the case in practice).

[2] These results are based on yield curve data from Gurkaynak et al. (2007) as described in Appendix B. We thank the authors for sharing these data.

[3] In Appendix B we provide full details on the data sources in the paper.

shape to those estimated from maturity-ordered financial assets. This is preliminary evidence that the high explained variation and the particular shape of PC loadings are reflective of the strong local correlation (i.e., smooth curve) across these ordered data. We will show that this behavior can be consistent with a much higher dimensional factor space.

Figure 1. Principal Component Loadings in Financial and Non-Financial Data. This figure presents the loadings of the first three principal components for bond yields (top left), oil price futures (top right), S&P 500 options returns (bottom left) and global surface temperature (bottom right). For each panel, principal components analysis is applied to the sample correlation matrix. A full description of the data and data sources is available in Appendix B.



An unappealing property of bond yields is that they are very persistent. Recent work by Uhlig (2009) and Onatski and Wang (2021) argue that extracting principal components from factorless, nonstationary data results in a spurious inference of a small number of factors that absorb almost all of the variation in the data. To evade this spurious factor problem, the data should be transformed to induce stationarity such as using first (time-series) differences, $y_t^{(n)} - y_{t-1}^{(n)}$, or bond returns. In fact, this was precisely the approach taken in the seminal work of the literature (Litterman and Scheinkman (1991) and Garbade (1996)) on extracting principal components from bond data. Define the one-period holding return on a bond of maturity n from time t to $t+1$ as

$$r_{t,t+1}^{(n)} \equiv p_{t+1}^{(n-1)} - p_t^{(n)}. \tag{3}$$

The corresponding excess return is then

$$rx_{t,t+1}^{(n)} \equiv r_{t,t+1}^{(n)} - y_t^{(1)} = r_{t,t+1}^{(n)} - r_{t,t+1}^{(1)}. \tag{4}$$

The notation $r_{t,t+1}^{(n)}$ and $rx_{t,t+1}^{(n)}$ signifies that these returns are earned from period t to $t+1$. In the sequel, we will simplify notation to $r_{t+1}^{(n)}$ and $rx_{t+1}^{(n)}$, respectively. Observe that since returns are defined based on changes in prices, they have two important properties: (i) the (approximate) time differencing substantially reduces the degree of persistence; (ii) the overlapping nature of prices relative to forward rates is directly inherited by returns. We can demonstrate the second property in a straightforward way by defining the return on the following long/short trading strategy: buy an n-maturity bond and short an $(n-1)$-maturity bond ("difference return")

$$dr_{t+1}^{(n)} \equiv \left(p_{t+1}^{(n-1)} - p_t^{(n)}\right) - \left(p_{t+1}^{(n-2)} - p_t^{(n-1)}\right) = r_{t+1}^{(n)} - r_{t+1}^{(n-1)} = f_t^{(n)} - f_{t+1}^{(n-1)}. \tag{5}$$

Then, it follows immediately that excess returns are a partial sum (along maturity) of $dr_{t+1}^{(i)}$,

$$rx_{t+1}^{(n)} = \sum\nolimits_{i=2}^{n} dr_{t+1}^{(i)}. \tag{6}$$

If we stack excess returns and difference returns in the $(N-1)\times T$ matrices R and D, respectively, then equation (6) implies

$$R = C_1 D \quad \text{or} \quad D = C_1^{-1} R, \tag{7}$$

where C_1 is a lower triangular matrix of ones. Thus, difference returns and excess returns are related by a simple nonsingular transformation.

To illustrate the difficulties in determining the minimal dimension of the term structure, we use a simplified example from Crump and Gospodinov (2021). The covariance matrix of R is

$V_R = C_1 V_D C_1'$, where V_D is the covariance matrix of D. For simplicity, assume that $V_D \propto I_{N-1}$ so that $V_R = \sigma^2 C_1 C_1'$. Principal components analysis (PCA) is based on the eigendecomposition of V_R. This is a non-factor model with difference returns being the driving primitive process (of dimension $N-1$). However, the factor structure is extracted from excess returns that exhibit strong cross-sectional dependence due to the overlapping maturities in adjacent excess returns. To see this, note that by equation (6), excess returns of maturities n_1 and n_2 will have $\min(n_1, n_2) - 1$ differenced returns in common.

The factor loadings are the eigenvectors $\psi_j = (\psi_{1,j}, \ldots, \psi_{N-1,j})'$ of matrix V_R given by

$$\psi_{h,j} = \frac{2}{\sqrt{2N-1}} \sin\left(\frac{h(2j-1)\pi}{2N-1}\right). \tag{8}$$

for $h = 1, \ldots, N-1$ and $j = 1, \ldots, N-1$. Note that the shape of these factor loadings is fully characterized by only the order of the corresponding eigenvector and the time to maturity. Despite the simplicity and unrealistic nature of the design, Figure 2 shows that these analytical factor loadings mimic closely the highly structured polynomial pattern in the PC loadings for excess returns in the data.[4] In this example, the ordered eigenvalues of the V_R matrix can be expressed as

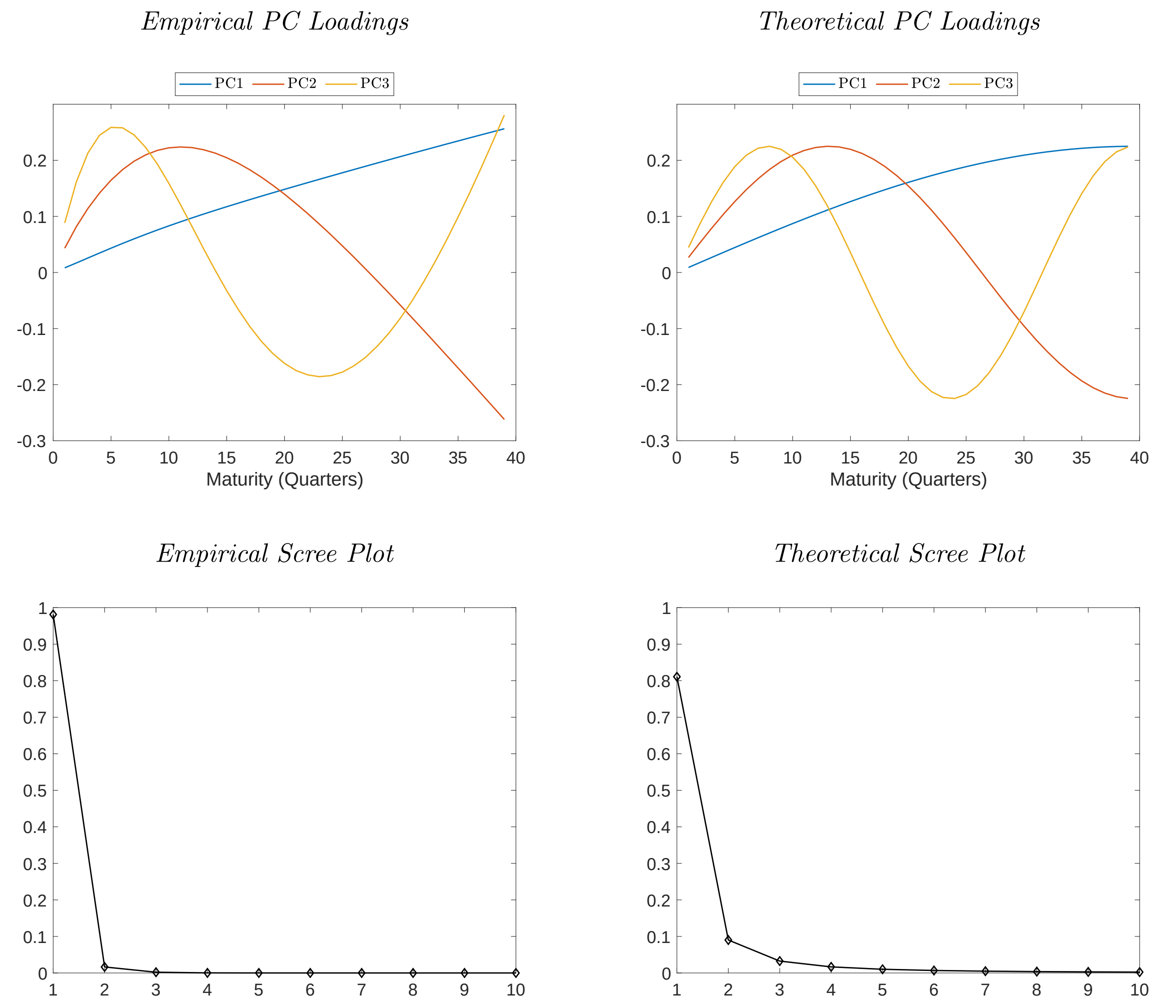
$$\lambda_j = \frac{\sigma^2}{2 - 2\cos\left(\frac{(2j-1)\pi}{2N-1}\right)}, \qquad \lambda_1 > \lambda_2 > \cdots > \lambda_{N-1}, \tag{9}$$

which, again, are only a function of time to maturity. Figure 2 shows that the first three eigenvalues account for more than 93% of the sum of the eigenvalues, $(N-1)N/2$.

The key takeaway from this example and the results in Figure 2 is that when $V_D \propto I_{N-1}$, there are $N-1$ idiosyncratic factors which are driving bond returns. However, the standard outputs from PCA suggests a much smaller dimension of the factor space. The intuition for this result is that the overlapping sum that links difference and excess returns, induces strong cross-sectional correlations which the PCA method interprets as evidence of only a small number of underlying factors. Crump and Gospodinov (2021) provide further evidence with more realistic setups, but the same conclusion applies; namely, that pinning down the dimension of the factor space is extremely

[4]The plots in the left column of Figure 2 use the same yield curve data from Gurkaynak et al. (2007) as the top left plot in Figure 1 except for excess returns rather than yields. Note that the difference in the shape of the loadings between Figures 1 and 2 is due to PCA applied to either the sample correlation matrix or the sample covariance matrix.

Figure 2. Empirical and Theoretical Principal Component Loadings of Bond Returns. The left column of this figure presents the first three principal component loadings for excess returns and the associated scree plot based on an eigendecomposition of the sample covariance matrix using data from Gurkaynak et al. (2007) for the sample 1972:Q1–2022:Q4. The right column shows the theoretical counterparts based on equations (8) and (9).



challenging.

However, the nature of the cross-sectional dependence is not the only uncertain characteristic of the true data-generating process. Estimation and inference is also affected by the time-series dynamics of the underlying series. We next show how these sources of uncertainty can hinder a purely parametric approach to inference, whereas our newly proposed bootstrap remains effective.

2.2 Illustrating Examples

2.2.1 Revisiting the "Tent Shape"

To motivate the need for a nonparametric bootstrap tailored to the term structure of interest rates, we revisit the work of Cochrane and Piazzesi (2005, 2008). Cochrane and Piazzesi (2005, 2008) investigate the relation between future one-year holding period returns and the current forward curve which is characterized by a "tent-shape" pattern in the estimated regression coefficients. Define $f_t^{(n,h)} \equiv p_t^{(n-h)} - p_t^{(n)}$ as the log-forward rate corresponding to a h-period investment between $t+n-h$ and $t+n$, and $rx_{t+h}^{(n,h)} \equiv p_{t+h}^{(n-h)} - p_t^{(n)} + p_t^{(h)}$ as the excess h-period holding return. We follow Cochrane and Piazzesi (2005) and run monthly predictive regressions of excess 12-month holding returns of the form

$$rx_{t+12}^{(n,12)} = \alpha^{(n)} + \beta_1^{(n)} \cdot y_t^{(12)} + \beta_2^{(n)} \cdot f_t^{(24,12)} + \beta_3^{(n)} \cdot f_t^{(36,12)} + \beta_4^{(n)} \cdot f_t^{(48,12)} + \beta_5^{(n)} \cdot f_t^{(60,12)} + v_t, \quad (10)$$

where $n \in \{24, 36, 48, 60\}$.

The black solid line in the left column of Figure 3 replicates the tent-shaped estimated regression coefficients averaged across the four maturities as in Cochrane and Piazzesi (2005) using Fama-Bliss discount bond data. The black solid line in the right column of the figure replicates the R^2 for each of the four regressions. A candidate resampling procedure should mimic these patterns that are observed in the data. The multi-colored lines in the top row of Figure 3 represent the same estimated quantities but from individual bootstrap replications based on the nonparametric bootstrap that we formally introduce in the next section. In the top left chart we can see that the bootstrapped estimates follow a similar contour as the original tent shape. Similarly, in the top right chart we observe that the pattern of R^2 is also similar showing a distinct kink at the 4-year maturity. It is important to emphasize that the nonparametric bootstrap does not take a stand on the factor structure or the exact time-series dynamics of the data. Despite this, these features in the data, based on estimates of equation (10), are preserved in the bootstrapped samples.

As a comparison we also implement the parametric bootstrap of Bauer and Hamilton (2018) which produces resampled yields based on an affine factor model using factors, estimated by principal components, that are assumed to follow a VAR(1). In the middle row we show the bootstrapped regression coefficients and R^2 based on a choice of 3 factors as originally proposed in Bauer and

Hamilton (2018). In contrast to the nonparametric bootstrap, the parametric alternative fails to capture the pronounced tent shape in the data. Moreover, the parametric bootstrap also misses key features of the regression R^2 and appears to be downward biased. As illustrated in the bottom row of Figure 3, saturating the factor space by shifting to 5 principal components cannot rectify these problems. Instead, the charts in the middle and bottom row are very similar. This is suggestive evidence that the parametric restriction on the time-series dynamics can also be an important source of misspecification.

2.2.2 Preliminary Simulation Evidence

We further investigate the properties of these bootstrapped methods using simulated data. Although the results shown thus far imply that we should be uneasy making strong parametric assumptions in the context of the yield curve, we will use these varied parametric-based Monte Carlo experiments to generate data and highlight the robustness of our bootstrap. Specifically, we will consider three different simulation designs. Informed by the application above, the factor structure of the term structure can be characterized by individual forward rates. We have,

$$F_t = \mathbf{a}^f + \mathbf{B}^f g_t + e_t, \tag{11}$$

where $F_t = (y_t^{(12)}, f_t^{(24,12)}, f_t^{(36,12)}, \ldots, f_t^{(108,12)}, f_t^{(120,12)})'$, g_t denotes a subset of forward rates that are assumed to follow a Gaussian vector autoregression (VAR), $\mathbf{a}^f$ and $\mathbf{B}^f$ are affine parameters, and e_t are independent (over time) Gaussian measurement error with a reduced-rank variance matrix.[5] Based on the simulated forwards, F_t, we can then construct simulated yields and returns (using equations (2), (3), and (4)), run regressions of the form

$$rx_{t+12}^{(n,12)} = \alpha^{(n)} + \beta^{(n)\prime} g_t + v_t, \tag{12}$$

and conduct inference on the elements of $\beta^{(n)}$. All our results are based on 1,000 simulations with a sample size $T = 600$. The values of the true parameters, $\beta^{(n)}$, can be obtained analytically and are given in Section 3.3. For power, we consider tests of $\mathbb{H}_0 : \beta_i^{(n)} = 0$, where $\beta_i^{(n)}$ is the ith element of $\beta^{(n)}$. We omit results for the constant term as it is not generally a parameter of interest in these

[5] All necessary parameters are calibrated using system OLS estimates based on Gurkaynak et al. (2007) data over the sample period 1972:m1–2022:m12.

Table 1. Simulations (VAR(1) Specification, 3 Forward Factors) This table presents empirical size and power for the nonparametric and parametric bootstrap methods described in the main text. The nominal level is 10% and the sample size is $T = 600$. Each column reports results for the t-test associated with the regressor $(g_{1t}, g_{2t}, g_{3t}) = (y_t^{(12)}, f_t^{(60,12)}, f_t^{(120,12)})$. Based on 1,000 simulations and 399 bootstrap replications per simulation.

	Nonparametric Bootstrap					
	Size			*Power*		
Maturity	g_{1t}	g_{2t}	g_{3t}	g_{1t}	g_{2t}	g_{3t}
2y	0.112	0.093	0.076	0.255	0.214	0.076
3y	0.099	0.110	0.082	0.415	0.335	0.090
4y	0.097	0.107	0.082	0.544	0.398	0.106
5y	0.102	0.102	0.090	0.617	0.419	0.110
6y	0.105	0.106	0.096	0.659	0.405	0.096
7y	0.107	0.103	0.102	0.687	0.368	0.096
8y	0.111	0.101	0.103	0.689	0.312	0.101
9y	0.113	0.098	0.102	0.683	0.258	0.134
10y	0.111	0.095	0.100	0.673	0.201	0.173
	Parametric Bootstrap					
	Size			*Power*		
Maturity	g_{1t}	g_{2t}	g_{3t}	g_{1t}	g_{2t}	g_{3t}
2y	0.055	0.033	0.031	0.253	0.161	0.017
3y	0.054	0.035	0.023	0.416	0.240	0.033
4y	0.051	0.028	0.005	0.530	0.319	0.028
5y	0.041	0.025	0.002	0.599	0.336	0.018
6y	0.040	0.016	0.001	0.638	0.330	0.001
7y	0.044	0.011	0.000	0.649	0.291	0.000
8y	0.042	0.008	0.000	0.648	0.232	0.000
9y	0.043	0.013	0.000	0.641	0.172	0.000
10y	0.043	0.010	0.000	0.632	0.094	0.000

applications.

In Tables 1 and 2, we report both the empirical size and power of tests based on our nonparametric bootstrap and the parametric approach exactly as detailed in Bauer and Hamilton (2018). For both bootstrap methods, we use $B = 399$ bootstrap replications. We also investigate the empirical size based on commonly-used heteroskedasticity and autocorrelation consistent/robust (HAC/HAR) variance estimators (Newey and West (1987) and Lazarus et al. (2018)) which rely on asymptotic approximations to conduct feasible inference. We relegate these results to the SA as they fail to control size uniformly across all of our simulation designs.

For the first specification, we choose $g_t = (y_t^{(12)}, f_t^{(60,12)}, f_t^{(120,12)})'$ and assume that g_t follows a VAR(1) with parameters calibrated from the data. This setup aligns closely with the parametric structure in Bauer and Hamilton (2018). Table 1 presents the results for this design. The empirical size for our nonparametric bootstrap is very close to the nominal size (10%) uniformly across maturities and for all three factors.[6] It is important to note that the nonparametric bootstrap is

[6] In the main text and SA, we report results for a nominal size of 10%. Results for a nominal size of 5% are

Table 2. Simulations (VAR(1) Specification, 5 Forward Factors) This table presents empirical size and power for the nonparametric and parametric bootstrap methods described in the main text. The nominal level is 10% and the sample size is $T = 600$. Each column reports results for the t-test associated with the regressor $(g_{1t}, g_{2t}, g_{3t}, g_{4t}, g_{5t}) = (y_t^{(12)}, f_t^{(36,12)}, f_t^{(60,12)}, f_t^{(84,12)}, f_t^{(120,12)})$. Based on 1,000 simulations and 399 bootstrap replications per simulation.

Nonparametric Bootstrap

	Size					*Power*				
Maturity	g_{1t}	g_{2t}	g_{3t}	g_{4t}	g_{5t}	g_{1t}	g_{2t}	g_{3t}	g_{4t}	g_{5t}
2y	0.097	0.081	0.099	0.089	0.093	0.353	0.511	0.560	0.570	0.488
3y	0.091	0.086	0.095	0.092	0.091	0.392	0.428	0.468	0.524	0.474
4y	0.086	0.087	0.102	0.092	0.091	0.384	0.352	0.400	0.494	0.483
5y	0.087	0.089	0.105	0.087	0.096	0.378	0.306	0.367	0.500	0.513
6y	0.093	0.089	0.108	0.091	0.104	0.380	0.288	0.359	0.515	0.535
7y	0.101	0.096	0.107	0.096	0.106	0.380	0.285	0.371	0.540	0.544
8y	0.101	0.100	0.107	0.095	0.100	0.392	0.284	0.388	0.549	0.546
9y	0.108	0.100	0.106	0.099	0.103	0.401	0.289	0.400	0.555	0.516
10y	0.111	0.104	0.110	0.098	0.110	0.405	0.292	0.404	0.544	0.478

Parametric Bootstrap

	Size					*Power*				
Maturity	g_{1t}	g_{2t}	g_{3t}	g_{4t}	g_{5t}	g_{1t}	g_{2t}	g_{3t}	g_{4t}	g_{5t}
2y	0.043	0.022	0.016	0.018	0.027	0.023	0.000	0.000	0.000	0.000
3y	0.031	0.009	0.014	0.011	0.022	0.082	0.000	0.000	0.000	0.000
4y	0.024	0.002	0.018	0.009	0.024	0.125	0.000	0.000	0.000	0.000
5y	0.020	0.005	0.013	0.010	0.025	0.125	0.000	0.000	0.000	0.000
6y	0.016	0.006	0.019	0.010	0.018	0.151	0.000	0.000	0.000	0.000
7y	0.017	0.005	0.018	0.009	0.019	0.154	0.000	0.000	0.000	0.000
8y	0.019	0.009	0.011	0.010	0.020	0.147	0.000	0.000	0.000	0.000
9y	0.024	0.014	0.011	0.009	0.019	0.131	0.000	0.000	0.000	0.000
10y	0.023	0.012	0.013	0.010	0.017	0.138	0.000	0.000	0.000	0.000

completely agnostic about how the term structure data are generated. By contrast, the parametric bootstrap of Bauer and Hamilton (2018) under-rejects relative to the nominal size and for some maturities this under-rejection is severe. This is despite the fact that both the dimension of the factor space and the time-series dynamics are correctly specified.

In terms of power, the rejection rates for the nonparametric bootstrap are relatively high for the first factor and generally decline for the second and third factor. This can be explained by the relative proximity of the true coefficients of the second and third factor to zero. Given the under-rejections for the correctly centered test statistic, the power for the parametric bootstrap is reasonably high for the first factor but meaningfully lower for the other two factors.

For the second design, we increase the number of factors to five by adding $f_t^{(36,12)}$ and $f_t^{(84,12)}$. This will induce a source of misspecification for the BH bootstrap as its prescription is to utilize only three principal components of yields. However, this misspecification would not be easily

qualitatively similar and omitted to conserve space. These results are available upon request.

detected as, averaging across simulations, the percentage of variation in yields explained by the first three principal components is 97.39%, 2.48%, and 0.12%, respectively. Table 2 presents the empirical size and power for this specification. Again, our nonparametric bootstrap has excellent size properties while the parametric bootstrap is undersized for all factors. The power of the parametric bootstrap is severely compromised by the misspecification due to a strong estimation bias induced by the incorrect rotation of the factors. Although there is some power to discriminate the coefficient associated with the first factor from zero, for all other factors, the test fails to reject across all simulations. The intuition for this result is that, by bootstrapping based on a factor space of dimension three, the bootstrapped second moments fail to reproduce their population counterparts.

Another source of misspecification can arise from the time-series dynamics of the factors. In our third specification, we consider again a three-factor model but now with factors following a VAR(2). The results follow a very similar pattern as in Table 2 and are presented in the SA (see Table SA.4).

In summary, it is preferable to avoid committing to a tightly parameterized, finite-dimensional factor structure as it may result in the omission of important information embedded in the yield curve. This insight informs our resampling approach.

3 Bootstrap Inference in Bond Return Predictive Regressions

In this section, we introduce our bootstrap procedure and study its theoretical and finite-sample properties. The motivation for the bootstrap is to improve inference in common empirical settings in the asset pricing literature such as predictive return regressions to test the expectations hypothesis or estimate risk premia, or perform robust inference in parametric models of the term structure. Our resampling procedure can also be used for generating conditional future paths of yields or other yield-related variables as well as measures of sampling uncertainty around projected paths. This can be employed, for example, in forecasting, policy analyses, or the computation of option-adjusted spreads.

Our proposed bootstrap procedure is model-free and remains agnostic about the exact factor structure in the data and the form of time-series dependence. We resample the primitive objects $dr_t^{(n)}$, augmented with the longest horizon forward rate $f_t^{(N)}$. The intuition for our bootstrap method can be gleaned from the following identity,

$$f_t^{(n)} = f_t^{(n)} + f_{t-1}^{(n+1)} - f_{t-1}^{(n+1)} + \cdots + f_{t-N+n}^{(N)} - f_{t-N+n}^{(N)} \tag{13}$$

$$= f_{t-N+n}^{(N)} + dr_t^{(n+1)} + dr_{t-1}^{(n+2)} + \cdots + dr_{t-N+n+1}^{(N)}. \tag{14}$$

In particular, the relation between forwards across time informs the relations between future returns and the current term structure. Equation (14) implies that if we observe the longest horizon forward rate, $\{f_t^{(N)}\}_{t=2}^T$, all difference returns, $\{(dr_t^{(2)}, \ldots, dr_t^{(N)})\}_{t=2}^T$, along with an initial forward curve, $(f_1^{(1)}, \ldots, f_1^{(N)})$, we are able to construct the entire forward curve $\{(f_t^{(1)}, \ldots, f_t^{(N)})\}_{t=1}^T$ and yield curve $\{(y_t^{(1)}, \ldots, y_t^{(N)})\}_{t=1}^T$.[7] Thus, conditional on an initial forward curve, we may jointly resample the N-maturity forward and the difference returns to generate bootstrap samples of the whole yield curve.[8] A key feature of our bootstrap procedure is that we have chosen primitive objects in such a way that the only persistent primitive object is the longest horizon forward rate which is confined to a single dimension and is also less persistent than other yields or forwards. Importantly, reconstructing the yield curve via these identities allows the bootstrap procedure to mimic the salient features, such as time-series and cross-sectional dependence, of the bond data that we observe in practice.

A natural question that might arise is why not resample $\{(f_t^{(1)}, \ldots, f_t^{(N)})\}_{t=1}^T$ directly? Although using forward rates avoids the mechanical cross-sectional dependence discussed in Section 2, the strong time-series dependence is retained across all maturities. Instead, by using difference returns we can avoid both strong cross-sectional and time-series dependence.

[7]For $t \leq N+1$, we can obtain forwards by directly relying on the recursive relationship, $f_t^{(n)} = f_{t-1}^{(n+1)} - dr_t^{(n+1)}$.

[8]Observe that by equation (6) we can equivalently think of the bootstrap as resampling $\{(rx_t^{(2)}, \ldots, rx_t^{(N)})\}_{t=2}^T$ instead of the difference returns, $\{(dr_t^{(2)}, \ldots, dr_t^{(N)})\}_{t=2}^T$.

3.1 Bootstrapping the Yield Curve

Assume that $f_1^{(1)}, f_1^{(2)}, ..., f_1^{(N)}$ are given. Next, we stack $f_{t+1}^{(N)}$ and $dr_{t+1}^{(n+1)}$, $n = 1, ..., N-1$ and $t = 1, ..., T-1$, into the matrix

$$Z = \begin{bmatrix} f_2^{(N)} & dr_2^{(2)} & dr_2^{(3)} & ... & dr_2^{(N)} \\ f_3^{(N)} & dr_3^{(2)} & dr_3^{(3)} & ... & dr_3^{(N)} \\ ... & ... & ... & ... & ... \\ f_t^{(N)} & dr_t^{(2)} & dr_t^{(3)} & ... & dr_t^{(N)} \\ ... & ... & ... & ... & ... \\ f_T^{(N)} & dr_T^{(2)} & dr_T^{(3)} & ... & dr_T^{(N)} \end{bmatrix}.$$

We impose regularity conditions (stationarity and ergodicity) on the multivariate process $z_t = (f_t^{(N)}, dr_t^{(2)}, ..., dr_t^{(N)})'$ that would guarantee validity of the block bootstrap. The bootstrap samples $\{z_t^*\}$ for $t = 2, ..., T$ are then obtained by drawing with replacement blocks of $M = M_T \in \mathbb{Z}_+$ $(1 \leq M < T)$ rows (time-series dimension) from the matrix Z, jointly for all cross-sectional observations. This resampling structure allows for unknown forms of (possibly strong) cross-sectional dependence. Importantly, while the bootstrap can capture and preserve a strong common factor structure in the data, it remains agnostic about the precise source of cross-sectional dependence. It also deals with general forms of serial correlation provided that the time-series dependence is of mixing type. Furthermore, since the bootstrap is model-free, it is robust to possible model misspecification.

Let z_t be the t-th row of the data matrix Z above. Also, let $Z_{t,M} = (z_t, z_{t+1}, ..., z_{t+M-1})$ denote a block of M consecutive observations of z_t, $k = [T/M]$, where $[a]$ signifies the largest integer that is less than or equal to a, and $\overline{T} = kM$. We resample with replacement k blocks from ($Z_{1,M}$, $Z_{2,M}$, ..., $Z_{\overline{T}-M+1,M}$) by drawing k *iid* uniform random variables $[u_1], ..., [u_k]$ on $(1, k+1)$. Then, the bootstrap sample is given by $Z^* = [(z_1^*,\ z_2^*,\ ...,\ z_M^*),\ (z_{M+1}^*,\ z_{M+2}^*,\ ...,\ z_{2M}^*),\ ...,\ (z_{\overline{T}-M}^*,\ z_{\overline{T}-M+1}^*,\ ...,\ z_{\overline{T}}^*)] = (Z_{[u_1],M},\ Z_{[u_2],M},\ ...,\ Z_{[u_k],M})$. To induce stationarity in the bootstrap world, we use the circular block bootstrap that "wraps" the data (Politis and Romano (1994)). This is intended to rectify the heterogeneous nature of the distribution of z_t^*. For a given block size M, we then construct the blocks $\{Z_{t,M}\}_{t=1}^{\overline{T}/M}$.

Given the bootstrap sample $\{f_t^{(N)*}, dr_t^{(2)*}, ..., dr_t^{(N)*}\}$ and the initial forward curve,[9] the boot-

[9]In all numerical results we initialize the forward curve with the first sample observation. However, there may

strap forward rates with maturities 1 to $N-1$ periods are constructed for $t = 1, ..., T-1$ as

$$f_{t+1}^{(N-1)*} = f_t^{(N)*} - dr_{t+1}^{(N)*} \tag{15}$$

$$f_{t+1}^{(N-2)*} = f_t^{(N-1)*} - dr_{t+1}^{(N-1)*} \tag{16}$$

$$\vdots$$

$$f_{t+1}^{(1)*} = f_t^{(2)*} - dr_{t+1}^{(2)*}. \tag{17}$$

Finally, the bootstrap returns and yields are obtained, respectively, as $rx_t^{(n)*} = \sum_{i=2}^{n} dr_t^{(i)*}$ and $y_t^{(n)*} = \frac{1}{n}\sum_{i=1}^{n} f_t^{(i)*}$ for $n = 1, ..., N$ and $t = 2, ..., T$.[10]

Because the resampled Z^* could be used for simulating the entire yield curve, it is desirable to establish first that the bootstrap approximates accurately the key moments of the true distribution. Given the panel structure of the data, the conditions for bootstrap validity follow closely those in Goncalves and White (2002) and Goncalves (2011). These conditions are collected in Assumption A1 in Appendix A.1. Since in term structure data the cross-sectional dimension N is generally a nontrivial fraction of the sample size T, it is convenient to allow N to be a function of T (see also Valkanov (1998) for a similar parameterization). Moreover, this informs our choice of M which is guided by a simple rule-of-thumb setting $M = (TN)^{2/5}$.

In what follows, P^* denotes the probability measure induced by the bootstrap resampling, conditional on the data. Also, for a sequence of bootstrap statistics Z^*_{NT}, $Z^*_{NT} \xrightarrow{P^*} 0$ in probability signifies that for any $\epsilon > 0, \delta > 0$, $\lim_{N,T\to\infty} P[P^*(|Z^*_{NT}| > \delta) > \epsilon] = 0$. Finally, let $V_{NT} = \frac{1}{NT}\sum_{n=1}^{N}\sum_{t=1}^{T} E[(z_{nt} - E(z_{nt}))(z_{nt} - E(z_{nt}))']$.

Lemma 1. *Under Assumption A1, (a)* $\frac{1}{NT}\sum_{n=1}^{N}\sum_{t=1}^{T}(z^*_{nt} - z_{nt}) \xrightarrow{P^*} 0$ *in probability, and (b)* $\lim_{N,T\to\infty} P[P^*(|\hat{V}^*_{NT} - V_{NT}| > \delta) > \epsilon] = 0$ *for any* $\epsilon > 0, \delta > 0$.

The result in Lemma 1 is to demonstrate that the proposed model-free bootstrap approximates well the first two moments of the true distribution of the data. This is useful when bootstrapping the unconditional distribution of the entire yield curve.

be circumstances where it is desirable to conduct inference conditional on a different initial condition. Our results accommodate such a case.

[10]Since $f_t^{(N)}$ is typically a persistent process, we discuss below the option of prewhitening $f_t^{(N)}$ (and any other persistent variables) via an approximate AR(1) or VAR(1) model. See Remark 3 below.

To visualize how well our resampling procedure mimics the actual dynamics of various yield curve-based variables, Figure 4 plots a typical bootstrap realization against the observed data. It is reassuring to see that not only are the time-series dynamics of all these variables quite reasonable, but also that the bootstrap preserves and replicates the cross-sectional dependence across maturities.[11]

Remark 1. In some cases, data sets of bond prices do not have a short-rate ($n = 1$) available or begin at some maturity n_0 where $n_0 > 2$. For example, the German Bund yield curve published by the Bundesbank does not have a short-rate directly available (see Speck (2021) for further discussion). Alternatively, depending on the application at hand, it might be desirable to only resample a specific segment of the maturity spectrum. In this case, we can straightforwardly modify the recursive relationships in equations (15)–(17) for this setting. In particular, we can simply replace the terminal recursion in equation (17) by

$$f_{t+1}^{(n_0)*} = f_t^{(n_0+1)*} - dr_{t+1}^{(n_0+1)*},$$

and generate bootstrap samples for $\left\{ \left(y_t^{(n_0)}, y_t^{(n_0+1)}, \ldots, y_t^{(N)} \right) \right\}_{t=1}^{T}$. $\square$

Remark 2. Some data sets of bond prices do not include a sequence of adjacent maturities up to N. The most prominent example is the Fama-Bliss discount bond data, used in Section 2.2, which are available at monthly frequency in the time series but only annual data in the cross-sectional dimension. In this case, we can follow a similar design to the procedure described above. When the data are monthly observations with annual maturities, we observe $\left\{ \left(p_t^{(12)}, p_t^{(24)}, p_t^{(36)}, p_t^{(48)}, p_t^{(60)} \right) \right\}_{t=1}^{T}$. Then, the analog of $f_t^{(N)}$ in this setting is

$$z_{1t} = \left(p_t^{(48)} - p_t^{(60)} \right).$$

Similarly, the analog to the difference returns in this setting is

$$\begin{aligned} z_{2t} &= p_t^{(36)} - p_t^{(48)} - \left(p_{t-1}^{(48)} - p_{t-1}^{(60)} \right) \\ z_{3t} &= p_t^{(24)} - p_t^{(36)} - \left(p_{t-1}^{(36)} - p_{t-1}^{(48)} \right) \end{aligned}$$

[11] Our bootstrap can be further modified by adjusting for yield level dependence of volatility (see Rebonato and Zanetti (2023)) but we do not pursue this possibility further in this paper.

$$
\begin{aligned}
z_{4t} &= p_t^{(12)} - p_t^{(24)} - \left(p_{t-1}^{(24)} - p_{t-1}^{(36)}\right) \\
z_{5t} &= p_t^{(0)} - p_t^{(12)} - \left(p_{t-1}^{(12)} - p_{t-1}^{(24)}\right) = -p_t^{(12)} - \left(p_{t-1}^{(12)} - p_{t-1}^{(24)}\right).
\end{aligned}
$$

We can bootstrap $(z_{1t}, z_{2t}, \ldots, z_{5t})'$, conditional on the first observation, to obtain $\{(z_{1t}^*, z_{2t}^*, \ldots, z_{5t}^*)\}_{t=2}^T$ recursively, e.g.,

$$
p_2^{(36)*} - p_2^{(48)*} = z_{22}^* + \left(p_1^{(48)} - p_1^{(60)}\right),
$$

and so on. □

3.2 Predictive Regression Inference

In addition to simulating the yield curve, we are often interested in testing hypotheses related to the term structure of interest rates: testing the expectations hypothesis, running predictive return regressions to estimate or study risk premia, etc. For example, suppose that interest lies in a predictive regression of an individual excess return, $rx_{t+h}^{(n,h)}$ on a k-vector of predictors x_t, where the predictor vector x_t is partitioned as $x_t = (g_t', w_t')'$ with g_t denoting yield-based predictors and w_t being macro or other financial predictors. More explicitly, we focus on inference in the following predictive regression model

$$
\begin{aligned}
rx_{t+h}^{(n,h)} &= \alpha^{(n)} + \beta^{(n)\prime} x_t + \varepsilon_{t+h}^{(n)} && (18) \\
&= \alpha^{(n)} + \beta_1^{(n)\prime} g_t + \beta_2^{(n)\prime} w_t + \varepsilon_{t+h}^{(n)}, && (19)
\end{aligned}
$$

where $\varepsilon_{t+h}^{(n)}$ denote predictive regression errors. Following Cochrane and Piazzesi (2005), we also consider a cross-sectional average of bond returns, $\overline{rx}_{t+h} = \frac{1}{\#N_s}\sum_{n\in N_s} rx_{t+h}^{(n,h)}$, $N_s \subseteq \{2, \ldots, N\}$, by replacing $rx_{t+h}^{(n,h)}$ with $\overline{rx}_{t+h}$ in the above model.

The vector of yield-based predictors may include a long-maturity interest rate or forward rate, the spread between a long-maturity and short maturity yield or forward (e.g., Fama and Bliss (1987), Campbell and Shiller (1991)), a linear combination of yields (e.g., Joslin et al. (2011)) or forwards (e.g., Cochrane and Piazzesi (2005, 2008)).

We augment the matrix Z with w_t and denote the new matrix by $\tilde{Z}$ so that its t-th row is given by $\tilde{z}_t = (f_t^{(N)}, dr_t^{(2)}, \ldots, dr_t^{(N)}, w_t')'$. This matrix is resampled using a moving block bootstrap as described above to obtain $\tilde{Z}^*$. The bootstrap sample $\{f_t^{(N)*}, dr_t^{(2)*}, ..., dr_t^{(N)*}\}$ is used for reconstructing forwards $f_t^{(n)*}$, yields $y_t^{(n)*}$ and bond returns $rx_t^{(n)*}$ and $rx_t^{(n,h)*}$. The bootstrap yield-based predictors g_t^* are formed from the resampled $y_t^{(n)*}$ which guarantees that all yield curve identities between variables hold in each bootstrapped sample. This internal consistency is vital to ensure that bootstrap-based estimation and inference can successfully approximate the stochastic behavior of the particular object of interest in each application.

The bootstrapped external predictors W^* are the last columns of $\tilde{Z}^*$, constructed as $W^* = [(w_1^*, w_2^*, ..., w_M^*), (w_{M+1}^*, w_{M+2}^*, ..., w_{2M}^*), ..., (w_{T-M}^*, w_{T-M+1}^*, ..., w_T^*)]$. The bootstrap OLS estimator in the predictive regression is given by

$$\hat{\beta}^{(n)*} = \left(\sum_{t=1}^{T-h}(x_t^* - \bar{x}^*)(x_t^* - \bar{x}^*)'\right)^{-1}\left(\sum_{t=1}^{T-h}(x_t^* - \bar{x}^*)(rx_{t+h}^{(n,h)*} - \overline{rx}_{t+h}^{(n,h)*})'\right),$$

where $\bar{x}^* = (T-h)^{-1}\sum_{t=1}^{T-h} x_t^*$ and $\overline{rx}_{t+h}^{(n,h)*} = (T-h)^{-1}\sum_{t=1}^{T-h} rx_{t+h}^{(n,h)*}$.

To show that the quantiles of the bootstrap distribution of $\sqrt{T}(\hat{\beta}^{(n)*} - \hat{\beta}^{(n)})$ approximate well, in some metric, the quantiles of $\sqrt{T}(\hat{\beta}^{(n)} - \beta^{(n)})$, we need to strengthen Assumption A1 and expand it to include $\varepsilon_{t+h}^{(n)}$ and x_t. This is done in Assumption A2 in Appendix A.1. Furthermore, Assumption A3 in Appendix A.1 provides sufficient conditions for HAC estimation of the variance matrices of $\hat{\beta}^{(n)}$ and $\hat{\beta}^{(n)*}$: $\hat{\Omega}_T$ and $\hat{\Omega}_T^*$, respectively. More specifically, $\hat{\Omega}_T = \hat{A}_T^{-1}\hat{B}_T\hat{A}_T^{-1}$, where $\hat{A}_T^{-1} = T^{-1}\sum_{t=1}^{T-h}(x_t - \bar{x})(x_t - \bar{x})'$ and $\hat{B}_T$ is the HAC estimator of the variance matrix of $(x_t - \bar{x})\hat{\varepsilon}_{t+h}^{(n)}$ (see the Appendix). The bootstrap analog $\hat{\Omega}_T^* = \hat{A}_T^{*-1}\hat{B}_T^*\hat{A}_T^{*-1}$ is obtained by using the bootstrapped predictors x_t^* and residuals $\hat{\varepsilon}_{t+h}^{(n)*}$ in these expressions. With this notation, define the t-statistics $\hat{t}_i = (\hat{\beta}_i^{(n)} - \beta_i^{(n)})/\mathrm{se}(\hat{\beta}_i^{(n)})$, where $\mathrm{se}(\hat{\beta}^{(n)}) = \sqrt{\mathrm{diag}(\hat{\Omega}_T)}$, and $t_i^* = (\hat{\beta}_i^{(n)*} - \hat{\beta}_i^{(n)})/\mathrm{se}(\hat{\beta}_i^*)$, where $\mathrm{se}(\hat{\beta}^{(n)*}) = \sqrt{\mathrm{diag}(\hat{\Omega}_T^*)}$. The next theorem establishes the validity of the bootstrap method.

Theorem 1. *Under Assumptions A1, A2, and A3, we have that for any $\epsilon > 0$,*

$$P\left(\sup_{x\in\mathbb{R}^k}\left|P^*\left(\sqrt{T}(\hat{\beta}^{(n)*} - \hat{\beta}^{(n)}) \leq x\right) - P\left(\sqrt{T}(\hat{\beta}^{(n)} - \beta^{(n)}) \leq x\right)\right| > \epsilon\right) \to 0$$

where $\leq$ applies to each component of the relevant vector and

$$P\left(\sup_{x\in\mathbb{R}}\left|P^*\left(t_i^* \leq x\right) - P\left(\hat{t}_i \leq x\right)\right| > \epsilon\right) \to 0$$

for $i = 1, \dots, k$.

PROOF. See Appendix A.

The result in Theorem 1 allows the construction of bootstrap confidence intervals for $\beta_i^{(n)}$ with asymptotically correct coverage probabilities. We construct $100(1-\alpha)\%$ confidence intervals for $\beta_i^{(n)}$ using the symmetric percentile-t method, based on B bootstrap replications, defined as

$$C_\alpha(\beta_i^{(n)}) = \left[\hat{\beta}_i^{(n)} - se(\hat{\beta}_i^{(n)}) q_t^* (1-\alpha) \, , \; \hat{\beta}_i^{(n)} + se(\hat{\beta}_i^{(n)}) q_t^* (1-\alpha)\right],$$

where $q_t^*(1-\alpha)$ denotes the $(1-\alpha)$-th quantile of the bootstrap distribution of t_i^*. The p-values for the symmetric percentile-t bootstrap method, that we report in the empirical section below, are obtained as

$$p_{|t_i|}^* = \frac{1}{B} \sum_{j=1}^{B} I\left\{|t_{i,j}^*| > |\hat{t}_i|\right\}, \tag{20}$$

where $I\{\cdot\}$ denotes the indicator function and $t_{i,j}^*$ is t_i^* for the jth bootstrapped sample.

Remark 3. While the long forwards tend to exhibit less persistence than the short rates, $f_t^{(N)}$ can still be very persistent depending on the frequency of the data and the time period. Similarly, some external predictors may also exhibit strong serial correlation. At the very least, it may be convenient to transform the data in such a way that all variables in the matrix $\tilde{Z}$ are characterized by similar degree of time-series dependence. In this case, we "pre-whiten" jointly the processes $f_t^{(N)}$ and w_t using a VAR(1) model (or an AR(1) for $f_t^{(N)}$ when there are no additional variables, w_t) for approximating their dynamics and perform the block bootstrap on the residuals from this model (see Remark 7). This is a version of the hybrid bootstrap of Davison and Hinkley (1997) and Niebuhr et al. (2017) which combines the autoregressive and nonparametric block bootstrap approaches. It is important to underscore that the AR(1) or VAR(1) models are likely to be misspecified and they are intended to only weaken the underlying persistence so that any remaining time-series dependence is handled by the block bootstrap resampling. To outline the main steps in this hybrid bootstrap, let $\tilde{f}_t^{(N)}$ and $\tilde{w}_t$ denote the pre-whitened processes that are the residuals from fitting a joint VAR(1) model of $f_t^{(N)}$ and w_t with projection coefficient matrix $\hat{\Psi}$.[12] These pre-whitened

[12] In all empirical applications and simulation experiments we use the bootstrap bias correction as suggested by

processes are used in the matrix $\tilde{Z}$ which is resampled by the block bootstrap to obtain their bootstrap versions $\tilde{f}_t^{(N)*}$ and $\tilde{w}_t^*$. The bootstrap series $\tilde{f}_t^{(N)*}$ and $\tilde{w}_t^*$ are in turn used to construct ("re-whiten") the original predictors recursively as

$$\begin{pmatrix} f_t^{(N)*} \\ w_t^* \end{pmatrix} = \hat{\mu} + \hat{\Psi} \begin{pmatrix} f_{t-1}^{(N)*} \\ w_{t-1}^* \end{pmatrix} + \begin{pmatrix} \tilde{f}_{t-1}^{(N)*} \\ \tilde{w}_{t-1}^* \end{pmatrix} \tag{21}$$

for $t = 2, ..., T$, $f_1^{(N)*} = f_1^{(N)}$ and $w_1^* = w_1$. We then proceed as above. Note that the regularity conditions in Assumptions A1 and A2 in Appendix A.1 need to be strengthened to ensure the validity of this hybrid bootstrap procedure. For details, see Niebuhr et al. (2017). □

Remark 4. The h-period predictive setup in our paper implies a particular (quasi-overlapping) structure for the regression errors which is known to be poorly approximated by the standard HAC estimators. A possible alternative is to utilize larger-than-usual bandwidth/truncation parameter in the HAC estimation in an effort to reduce the size distortions of the off-the-shelf HAC variance estimators (Lazarus et al. (2018)). Our setting appears to feature too much time-series persistence for the effectiveness of such an approach. Another possibility is to resort to the percentile bootstrap method which does not require an estimate of the variance matrix. Instead, we rely on the moving block bootstrap to correct the size distortions induced by the HAC estimator. For example, Goncalves and Vogelsang (2011) show that this "naive" block bootstrap achieves the same first-order accuracy as the fixed-b asymptotic distribution (Kiefer and Vogelsang (2005)) of the original statistic. Having said that, we acknowledge that the complex serial correlation structure of the errors and the potential high persistence of the predictors call for a careful approach to choosing kernel shape and truncation parameters in this setup. □

Remark 5. While some yield-based predictors (yields, forwards, spreads) are possibly persistent processes, they are generated via identities of the primitive objects that we bootstrap.[13] As a result, the persistence of these predictors arises naturally from the maturity relationships embedded in the term structure of interest rates. Thus, the bootstrap is expected to approximate well the

Kilian (1998). Similar bootstrap-based bias correction has been utilized in the context of term structure applications before. See for example, Bauer et al. (2012) and Bauer et al. (2014).

[13]Recall that the dependent variable $rx_{t+h}^{(n,h)}$ and the predictors g_t are both obtained from the same underlying data, $y_t^{(n)}$ for $n = 1, ..., N$. However, conventional resampling of $rx_{t+h}^{(n,h)}$ and g_t within a regression framework would violate the definitional relationships described in Section 2.

finite-sample distribution of the estimates corresponding to these persistent predictors although alternative representations (such as lag-augmentation; see, for example, Olea and Plagborg-Møller (2021)) may further insulate the inference method from the persistent properties of the predictors. □

Remark 6. The coverage rates of the bootstrap can be further improved by an equal-tailed double bootstrap. Suppose the first-stage bootstrap uses B_1 bootstrap replications and for a given $\beta_i^{(n)}$ ($i = 1, ..., k$), let $C^*_{\lambda_i}(\beta_i^{(n)}) = \left[\hat{\beta}^{(n)*}_{i,(1-\lambda_i)},\ \hat{\beta}^{(n)*}_{i,(\lambda_i)}\right]$ for some $1/2 < \lambda_i < 1$, where $\hat{\beta}^{(n)*}_{i,(\lambda)}$ is the λth percentile of the bootstrap distribution of $\hat{\beta}_i^{(n)}$. For each first-stage bootstrap replication j ($j = 1, ..., B_1$) and using the same resampling procedure, generate a vector of B_2 estimates $\hat{\beta}^{(n)**}_{i,j} = (\hat{\beta}^{(n)**}_{i,j,1}, ..., \hat{\beta}^{(n)**}_{i,j,B_2})$ and construct the corresponding interval $C^{**}_{\lambda_i,j}(\beta_i^{(n)}) = \left[\hat{\beta}^{(n)**}_{i,j,(1-\lambda_i)},\ \hat{\beta}^{(n)**}_{i,j,(\lambda_i)}\right]$. Then, solve for

$$\hat{\lambda}_i = \arg\min_{\lambda_i} \frac{1}{B_1} \sum_{j=1}^{B_1} I\left\{\hat{\beta}_i^{(n)} \in C^{**}_{\lambda_i,j}(\beta_i^{(n)})\right\} > 1 - \alpha.$$

The adjusted double bootstrap confidence interval is then constructed as $C^*_{\hat{\lambda}_i}(\beta_i^{(n)}) = \left[\hat{\beta}^{(n)*}_{i,(1-\hat{\lambda}_i)},\ \hat{\beta}^{(n)*}_{i,\hat{\lambda}_i}\right]$ (see McCarthy et al. (2018)). This is repeated for all $\beta_i^{(n)}$ ($i = 1, ..., k$). □

The advantages of our proposed bootstrap method can be better appreciated by pointing out some of the empirical regularities of bond prices (and their transformations) that make mimicking the original data so challenging. First, there is high time-series and cross-sectional persistence in yields. While transformations of bond yields, such as bond returns, remove the strong serial correlation in the yield data, they are still strongly cross-sectionally correlated.

A bigger challenge, in our view, for statistically modeling and bootstrapping yields directly is their extreme cross-sectional dependence. Part of this dependence is mechanistic (arising from telescoping sums and averages of primitive processes) but so strong that it could overwhelm and obscure the relevant information in the primitive objects. Instead, our bootstrap procedure operates on the primitive objects and then recovers the time-series and cross-sectional characteristics of the original data. The consequences of this strong cross-sectional dependence in yields and bond returns are illustrated analytically in Crump and Gospodinov (2021). Crump and Gospodinov (2021) also argue that characterizing the true factor space in the term structure of interest rates is quite challenging and committing to a low-dimensional parametric structure can result in large hedging and portfolio allocation errors. In contrast, our resampling method retains the underlying factor structure in the primitive processes and reconstructs the entire yield curve in a model-free, identity-preserving fashion.

Finally, as argued above, bond returns exhibit substantial predictability using past yield and forward rate information. A standard model-based bootstrap procedure faces a number of difficulties in bootstrapping data in an internally consistent fashion that will replicate this predictability as well as the other time-series and cross-sectional properties of the yield data. Our proposed bootstrap method provides this internal consistency in mimicking the salient properties of the data by accommodating unknown forms of factor structure, time-series persistence and cross-sectional dependence.

3.3 Further Simulation Evidence

In this section, we provide further simulation evidence based on return predictability regressions featuring yield-based factors and external predictors as in equation (19). We use the same maturities (1-year to 10-year) as in Section 2.2.2 and $h = 12$. We assume that the yield-based factors follow a Gaussian VAR(1),

$$g_t = \mu_g + \Psi_g g_{t-1} + \nu_{g,t}, \qquad t = 1, \ldots, T. \tag{22}$$

We choose g_t to be the one-year, 5-year and 10-year bond yields and calibrate the necessary parameters using system OLS estimates based on GSW data over the sample 1972:m1–2022:m12.

The external predictors also follow a Gaussian VAR(1),

$$w_t = \mu_w + \Psi_w w_{t-1} + \nu_{w,t}, \qquad t = 1, \ldots, T. \tag{23}$$

Thus, we have that the true coefficients associated with the variables w_t, $\beta_2^{(n)}$, are identically zero. The true value of the coefficients associated with g_t is

$$\beta_1^{(n)} = C_1 \left(\Xi_1 \mathbf{B}^f - \Xi_2 \mathbf{B}^f \sum_{i=0}^{h-1} \Psi_g^h \mu_g \right), \tag{24}$$

where Ξ_1 and Ξ_2 can be obtained as the conformable identity matrix with the first row removed and last row removed, respectively. As before, C_1 is a conformable lower triangular matrix of ones.

We choose two standard macroeconomic predictors: the 3-month percentage change in core CPI inflation and the 3-month percentage change in industrial production.[14] We then calibrate the necessary parameters for the VAR(1) model using system OLS estimates over the same sample period as for yields. We estimate the joint variance matrix of the innovations to the yield factors and the external predictors to replicate correlations observed in the data. Note that these predictors are quite persistent with the maximum eigenvalue of the autoregressive matrix equal to 0.93.

We generate $1,000$ replications for a sample size of $T = 600$. We report empirical size for the null hypothesis that each element in $\beta_2^{(n)}$ is equal to zero along with the (infeasible) empirical size for the three coefficients in $\beta_1^{(n)}$.

Our proposed bootstrap method is implemented as described in Section 3 in a fully nonparametric and data-driven way by block resampling the matrix Z and then reconstructing the yield curve $\left\{y_t^{(n)*}\right\}$ for $t = 1, ..., T$ and $n = 1, ..., N$. We pre-whiten $f_t^{(N)}$ and w_t jointly using a VAR(1) specification. We use the symmetric percentile-t bootstrap method to conduct inference based on Newey-West standard errors with lag length of h. Specifically, we use the resampled data to construct bootstrapped t-statistics for the coefficients of interest and construct p-values as discussed in Section 3 (equation (20)). We use the "rule of thumb" choice for the block size, M.

We compare the results using our bootstrap to the parametric bootstrap of Bauer and Hamilton (2018). We follow their implementation exactly. We use the first three principal components of yields along with system OLS estimates to generate bootstrap samples. The external predictors are bootstrapped under the null hypothesis of no return predictability. Note that this method is parametric in nature as it exploits the factor structure in yields, the specific parametric form of the dynamics, and proceeds under the null which holds in our design. In contrast, we emphasize that our resampling procedure is agnostic about how the simulated data were generated; we remain fully nonparametric. For both bootstrap methods, we use $B = 399$.

Table 3 presents results for inference on the coefficients $\beta_1^{(n)}$ and $\beta_2^{(n)}$. We first discuss the properties of the different inference procedures for the coefficients associated with the yield predictors. Our method controls size very well across maturities and the three different predictors. The parametric bootstrap, on the other hand, appears to be much more conservative with empirical size comfortably below the nominal size of 10%. For the external predictors, both methods work very well with empirical size very close to nominal size across maturities. We emphasize

[14]FRED mnemonics: `CPILFESL` and `INDPRO`, respectively.

Table 3. Simulations (VAR(1) Specification, 3 Yield Factors & 2 External Predictors) This table presents empirical size and power for the nonparametric and parametric bootstrap methods described in the main text. The nominal level is 10% and the sample size is $T = 600$. Each column reports results for the t-test associated with the regressor $(g_{1t}, g_{2t}, g_{3t}) = (y_t^{(12)}, y_t^{(60)}, y_t^{(120)})$. For the bivariate external regressors, w_t, only empirical size is reported. Based on 1,000 simulations and 399 bootstrap replications per simulation.

Nonparametric Bootstrap

	Size					*Power*		
Maturity	g_{1t}	g_{2t}	g_{3t}	w_{1t}	w_{2t}	g_{1t}	g_{2t}	g_{3t}
2y	0.088	0.073	0.068	0.085	0.089	0.137	0.065	0.141
3y	0.087	0.082	0.067	0.089	0.094	0.199	0.065	0.116
4y	0.079	0.079	0.065	0.092	0.097	0.232	0.071	0.118
5y	0.081	0.083	0.066	0.089	0.097	0.238	0.066	0.151
6y	0.087	0.083	0.069	0.080	0.098	0.238	0.067	0.206
7y	0.092	0.083	0.070	0.080	0.099	0.225	0.085	0.267
8y	0.094	0.083	0.068	0.085	0.102	0.209	0.110	0.339
9y	0.098	0.084	0.076	0.085	0.100	0.189	0.130	0.405
10y	0.097	0.084	0.076	0.089	0.105	0.165	0.151	0.478

Parametric Bootstrap

	Size					*Power*		
Maturity	g_{1t}	g_{2t}	g_{3t}	w_{1t}	w_{2t}	g_{1t}	g_{2t}	g_{3t}
2y	0.016	0.029	0.030	0.110	0.093	0.065	0.018	0.085
3y	0.014	0.033	0.031	0.102	0.096	0.119	0.020	0.056
4y	0.014	0.030	0.029	0.106	0.101	0.151	0.020	0.067
5y	0.012	0.028	0.028	0.105	0.107	0.171	0.018	0.107
6y	0.015	0.027	0.022	0.107	0.112	0.166	0.014	0.194
7y	0.014	0.022	0.023	0.108	0.110	0.144	0.023	0.256
8y	0.018	0.020	0.020	0.110	0.110	0.126	0.037	0.351
9y	0.018	0.018	0.025	0.106	0.110	0.106	0.069	0.445
10y	0.018	0.016	0.025	0.110	0.114	0.087	0.116	0.513

that the overlapping nature of these returns (for $h > 1$) produces strong serial correlation in the left-hand side variable.[15] With predictors which are also persistent, minimizing size distortion becomes increasingly difficult. Despite this challenging predictive regression design, both bootstrap procedures appear to provide a very accurate approximation to the finite-sample distribution of the t-statistics associated with $\beta_2^{(n)}$. In Section SA-1 of the SA we report the corresponding results based on HAC/HAR estimators which rely on asymptotic approximations to the distribution of the test statistic. These methods uniformly fail to control empirical size, especially for the external predictors.

Since the true value of $\beta_2^{(n)}$ is zero, we only report power for the yield predictors. In this design, discriminating the null hypothesis is challenging as it depends on the relative proximity of the true coefficients, that are calibrated to the actual data, to zero. That said, despite the relatively low power, the nonparametric bootstrap generally outperforms the parametric bootstrap.

[15] For a heuristic example, note that an h-period moving average of a white noise process has first-order autocorrelation coefficient of $\frac{h-1}{h}$.

4 Empirical Analyses

In this section, we introduce three empirical applications which highlight the versatility of our proposed bootstrap procedure. We first revisit regression-based tests of the expectations hypothesis and use our bootstrap to better understand the sampling properties of the corresponding estimators. We next investigate the predictive properties of trend inflation for future bond returns based on the work of Cieslak and Povala (2015). Finally, we construct bias-corrected estimates and confidence intervals for the probability of a future recession, based on information in the current yield curve.

4.1 Regression-Based Tests of the Expectations Hypothesis

Our methodological approach, which favors difference returns as the primitive object in a nonparametric bootstrap procedure, can be used to revisit existing analyses of the term structure of interest rates. Here, we investigate the properties of regression-based tests of the expectations hypothesis and introduce a new, alternative specification which is better suited to the realized properties of bond prices. Furthermore, because our bootstrap is not model specific, we can directly compare the statistical properties of the different specifications that have been used in the literature.

Our focus on difference returns as a primitive object allows us to unify and generalize the various regression-based tests of the expectations hypothesis (Fama and Bliss (1987), Campbell and Shiller (1991)). By the definition of forward rates and h-period returns, we have that

$$f_{t+m}^{(n-m)} = f_t^{(n)} - \left(rx_{t+m}^{(n,m)} - rx_{t+m}^{(n-1,m)} \right) \tag{25}$$

so that, under rational expectations, a regression-based implementation of a test of the expectations hypothesis can be based on the specification

$$f_{t+m}^{(n-m)} = \alpha + \beta \cdot f_t^{(n)} + \epsilon_{t+m}^{(n,m)}. \tag{26}$$

Following Fama and Bliss (1987), under the null hypothesis that $\beta = 1$, we can subtract the contemporaneous short rate from either side to obtain,

$$f_{t+m}^{(n-m)} - y_t^{(1)} = \alpha + \beta \cdot \left(f_t^{(n)} - y_t^{(1)} \right) + \epsilon_{t+m}^{(n,m)}. \tag{27}$$

We refer to this regression-based test of the expectations hypothesis as the CG test. Equation (27) forms the basis for all existing tests of the expectations hypothesis; specifically, under the null hypothesis, any existing test can be obtained through linear combinations of equation (27) for different n and m. To see this, note first that the test of Fama and Bliss (1987), denoted by FB, is implemented exactly as in equation (27) in the special case where $m = n - 1$, i.e.,

$$y_{t+n-1}^{(1)} - y_t^{(1)} = \alpha + \beta \cdot \left(f_t^{(n)} - y_t^{(1)} \right) + \epsilon_{t+n-1}^{(n,n-1)}. \tag{28}$$

Furthermore, a time-series average of equation (28) gives

$$\frac{1}{n} \sum_{s=1}^{n} y_{t+s-1}^{(1)} - y_t^{(1)} = \alpha + \beta \cdot \left(\frac{1}{n} \sum_{s=1}^{n} f_t^{(n)} - y_t^{(1)} \right) + \frac{1}{n} \sum_{s=1}^{n} \epsilon_{t+s-1}^{(s,s-1)}, \tag{29}$$

and using that $y_t^{(n)} = n^{-1} \sum_s f_t^{(s)}$, we obtain the first test of Campbell and Shiller (1991), denoted CS_1,

$$\frac{1}{n} \sum_{s=0}^{n-1} \left(y_{t+s}^{(1)} - y_t^{(1)} \right) = \alpha + \beta \left[y_t^{(n)} - y_t^{(1)} \right] + v_{\mathrm{CS}_1, t+n-1}^{(n)}, \tag{30}$$

where $v_{\mathrm{CS}_1,t+n-1}^{(n)} = \frac{1}{n} \sum_{s=1}^{n} \epsilon_{t+s-1}^{(s,s-1)}$. The second specification considered in Campbell and Shiller (1991), denoted CS_2, is

$$y_{t+m}^{(n-m)} - y_t^{(n)} = \alpha + \beta \frac{m}{n-m} \left(y_t^{(n)} - y_t^{(m)} \right) + v_{\mathrm{CS}_2, t+n-1}^{(n)}. \tag{31}$$

This can be obtained by taking a cross-sectional average of equation (27). To see this, note that

$$\frac{1}{n-m} \sum_{s=1}^{n-m} f_{t+m}^{(n-m)} = y_{t+m}^{(n-m)}, \qquad \frac{1}{n-m} \sum_{s=1}^{n-m} f_t^{(s+m)} = y_t^{(n)} + \frac{m}{n-m} \left(y_t^{(n)} - y_t^{(m)} \right),$$

and $v_{\mathrm{CS}_2,t+n-1}^{(n)} = \frac{1}{n-m} \sum_{s=1}^{n-m} \epsilon_{t+m}^{(s,m)}$.

Although FB, CS_1, CS_2 can all be derived from CG, the statistical properties may be very different. In fact, it has been well established in the literature that the results of tests of the expectations hypothesis can vary sharply depending on which regression specification is used (see Campbell (2017) for a comprehensive discussion). It is also well known that inference on the coefficient of interest in these regression-based tests is fraught with difficulties (see Bekaert et al. (1997) or Rossi

(2007), among others). In particular, statistical challenges arise from the high persistence of yields and forwards, the "unbalancedness" of some of the specifications,[16] the overlapping nature of some variables, and the unusual property that the dependent and explanatory variables are both derived from the same underlying yield curve. Our bootstrap is uniquely situated to better understand the statistical properties of the different versions of the regression-based tests of the expectations hypothesis. More specifically, we nonparametrically resample the entire yield curve, rather than the specific maturities used in a specification, which allows us to make comparisons across all of these formulations in a cohesive way. Any parametric resampling approach would necessarily be a model-specific approach which either imposed the expectations hypothesis or its failure.

To implement these tests, we use quarterly bond yield data from Gurkaynak et al. (2007) for maturities up to ten years for the sample period 1972Q1 to 2022Q4. Figure 5 provides the OLS estimates (black line) across different maturities for the four tests above, along with a horizontal line at $\beta = 1$ (dashed black line). For the CS_2 and CG tests, we require a choice of m. We present results for $m = 4$ (one year) and $m = 20$ (5 years). The top row shows the results for CS_2. For small m, we observe the well-known result that estimates of β are negative and large in magnitude providing strong counterfactual evidence against the expectations hypothesis.[17] When $m = 20$, the point estimate of β rises but is close to zero for most maturities. The results of 999 bootstrap draws are presented in grey.[18] Despite the point estimates being far away from one, in general, the bootstrapped OLS estimates are subject to a high degree of variability. In fact, the statistical evidence against the expectations hypothesis for the CS_2 test is generally weak with most bootstrap draws above the OLS estimates. Figure 6 sheds some light on why this might be the case. The CS_2 test features a highly unbalanced specification, especially when m is small. The top left chart of Figure 6 shows that when $m = 4$, the dependent and explanatory variables have strikingly different persistence and volatility properties. A key advantage of our bootstrap is that we do not resample these variables directly but instead we resample primitive objects with better statistical properties while still mimicking the important features of the data. Thus, the top row of Figure 5 provides a translation of the statistical problems, identified by the time-series plots, mapped to the sampling

[16] We define an unbalanced regression as a specification where the dependent and explanatory variables are characterized by meaningfully different degrees of persistence and variability.

[17] See, for example, Campbell (2017, p. 239): "...an estimated slope coefficient that not only fails to equal one but is actually negative."

[18] We deliberately choose the same scale for all graphs to demonstrate the different sampling behavior of the OLS estimator in each specification.

variability of the OLS estimator.

Conversely, the CS_1 and FB tests, in the middle row of Figure 5, have very different properties. First, the point estimates are much closer to one and are comfortably positive. Second, the variability – as judged by the bootstrap draws of the OLS estimate – is modulated substantially relative to the top row. However, the estimate of β appears unstable across maturities: a feature that our bootstrap mimics and is thus likely to be inherent to the test itself.

We can improve further upon all of these tests by using equation (27) directly. In the bottom row, we present results for $m = 4$ and $m = 20$ using CG. We observe that the OLS coefficient is substantially more stable across maturities. We also observe that the sampling uncertainty is lower than in the CS_1 and FB tests and significantly smaller than in the CS_2 test. We can link these results to the much more balanced dependent and explanatory variables as shown in Figure 6.

We now provide additional evidence by explicitly testing the null hypothesis that $\beta = 1$, i.e., a test of the expectations hypothesis. The CS_2 test for $m = 1$ is the most popular implementation in the literature and has been documented to provide the strongest evidence against the expectations hypothesis (for a recent example, see Farmer et al. 2022). This is performed using an asymptotic approximation to the t-statistic based on HAC/HAR standard errors. The top left chart in Figure 7 compares the p-values from this standard approach (labelled "Asymp. Approx.") as compared to those using the nonparametric bootstrap (labelled "Nonpar. Bootstrap"), for each n and $m = 1$. For the standard approach, we use the test statistic of Lazarus et al. (2018) with the equal-weighted cosine variance estimator. It is then straightforward to construct the corresponding p-values as the limiting distribution is Student's-t. Figure 7 shows that, in line with the existing literature, tests based on asymptotic distributional approximations overwhelmingly reject the expectations hypothesis across all maturities. In contrast, our bootstrap method provides a more mixed picture with most maturities featuring p-values above 5% and longer maturities well over 10%. In the right plot, we show the analogous p-values for the choice of $m = 4$. We observe that the gap between the two sets of p-values remains large although there is modestly more evidence against the expectations hypothesis in this specification. In the bottom two charts of Figure 7, we observe a similar pattern but for the CG test. The evidence against the expectations hypothesis is much stronger at shorter maturities than longer maturities. As mentioned above, an advantage of our nonparametric bootstrap is that we resample the entire yield curve in a uniform way which should better ensure consistency across different specifications. In particular, we come to similar conclusions – evidence is weakest at longer maturities – using either the CS_2 or CG bootstrap-based tests.

Overall, even with the improved sampling properties, our bootstrap finds only weak evidence against the expectations hypothesis. In the next section, we show that we can sharpen our bootstrap-based inference and find stronger evidence against the expectations hypothesis by alternative means using predictive return regressions with additional external predictors.

4.2 Trend Inflation and Bond Returns

The co-movement of inflation and nominal yields is a well established empirical fact in the United States and other economies. In addition, the dynamics of inflation have been shown to feature a time-varying, low-frequency component (e.g., Stock and Watson (2007)). Cieslak and Povala (2015) construct a measure of trend inflation, based on an exponentially smoothed moving average of year-over-year core consumer price index (CPI) inflation, and find that the inclusion of this variable substantially improves upon bond return predictability using only yields. However, the resulting trend inflation variable is highly persistent which complicates inference on its associated coefficient and the other regression coefficients in the model. Our proposed bootstrap is flexible enough to accommodate such a challenging setup, with yield *and* external predictors and very high persistence in at least one regressor. In fact, we can straightforwardly construct a resampling procedure that is tailored exactly to Cieslak and Povala (2015).

Cieslak and Povala (2015) run monthly regressions of the form given by equation (19) with $h = 12$ and

$$w_t = \frac{1-\phi}{1-\phi^{119}} \sum_{s=0}^{119} \phi^s \pi_{t-s}^{\text{yoy}}, \tag{32}$$

where π_t^{yoy} is the year-over-year growth rate in the core CPI and $\phi = 0.987$ is the choice of exponential smoothing parameter. We construct the necessary 12-month holding period returns for annual maturities from two years to ten years. As our choice for g_t, we use principal components of yields extracted from one-year to ten-year annual maturities. In the subsequent analysis, we use either the first two principal components (similar to the setup in Cieslak and Povala (2015)) or the first three principal components (as in Bauer and Hamilton (2018)).

To bootstrap the data in this setting, we follow the approach given in Section 3. We utilize a bivariate VAR(1) to accommodate the strong joint dynamics (as shown in the left plot of Figure 8),

$$\begin{pmatrix} f_t^{(N)} \\ \pi_t^{\text{yoy}} \end{pmatrix} = \mu + \Psi \begin{pmatrix} f_{t-1}^{(N)} \\ \pi_{t-1}^{\text{yoy}} \end{pmatrix} + v_t. \tag{33}$$

We estimate this VAR(1) via bias-corrected OLS (Kilian (1998)) and jointly block bootstrap $\hat{v}_t$ with the difference returns. This ensures that the bootstrapped data replicate the key features of the actual data. In the right plot of Figure 8, we show trend inflation (black line) along with bond yields (grey lines) for a particular bootstrap sample. We observe the strong co-movement in low-frequency dynamics between trend inflation and yields, just as in the actual data, which is shown in the left plot of Figure 8.

Also, because we require a ten-year burn-in period to construct w_t as in equation (32) we base our bootstrap on the sample starting ten years before the main sample. Then, for each bootstrap sample we obtain $\pi_t^{\text{yoy}*}$ and calculate w_t^* using equation (32). Finally, we drop the first 120 observations when constructing bootstrapped OLS coefficients and standard errors. By following these steps, we exactly mimic the steps being taken when estimating equation (19) on the actual data.

Table 4 presents the in-sample regression results using our nonparametric bootstrap for each individual maturity and the duration-weighted average return. For each of the two sample periods, 1971:m11–2022:m12 and 1983:m1–2022:m12, we report results using either two or three control variables. We omit the estimated coefficients on w_t for simplicity of presentation but note that they are all negative in sign and increasing in magnitude with maturity. For the full sample period and both specifications, we find that the bootstrap p-values associated with the coefficient of w_t are below 10% for bond returns based on maturities above five years. This suggests that the role of trend inflation for bond return predictability is most influential for longer maturity bonds.

For the shorter sample, 1983:m1–2022:m12, the bootstrap p-values are below 10% for all maturities and below 5% for the longest maturities. It appears that the results are weaker for the full sample because of the more complicated dynamics arising from multiple hump-shaped periods in the inflation process in the 1970s. Consequently, the bootstrap samples based on the longer sample exhibit more variability than those of the shorter sample which is dominated by the downward

Table 4. Bond Return Predictability of Trend Inflation This table presents bootstrap p-values for the coefficient of interest and 90% confidence intervals for R^2 based on the regression specification given by equation (19) with $h = 12$. The choice of w_t is the trend inflation estimate in equation (32) and the choice of g_t is either the first two (*2 Factors*) or the first three (*3 Factors*) principal components of yields. The sample period is 1971:m11–2022:m12 (full sample) or 1983:m1–2022:m12 (recent sample).

	Full Sample				**Recent Sample**			
	2 Factors		*3 Factors*		*2 Factors*		*3 Factors*	
Maturity	P-val	R^2 Conf. Int.	P-val	R^2 Conf. Int.	P-val	R^2 Conf. Int.	P-val	R^2 Conf. Int.
2y	0.228	[-0.001, 0.186]	0.198	[0.001, 0.251]	0.079	[0.000, 0.240]	0.086	[0.003, 0.322]
3y	0.189	[-0.001, 0.200]	0.168	[0.002, 0.271]	0.075	[0.001, 0.268]	0.076	[0.006, 0.346]
4y	0.151	[0.000, 0.218]	0.142	[0.003, 0.286]	0.071	[0.002, 0.291]	0.069	[0.010, 0.362]
5y	0.128	[0.000, 0.235]	0.122	[0.004, 0.300]	0.064	[0.004, 0.308]	0.062	[0.012, 0.379]
6y	0.108	[0.001, 0.247]	0.107	[0.006, 0.310]	0.060	[0.005, 0.324]	0.057	[0.015, 0.396]
7y	0.090	[0.002, 0.258]	0.096	[0.006, 0.318]	0.055	[0.007, 0.339]	0.052	[0.018, 0.410]
8y	0.077	[0.002, 0.267]	0.086	[0.008, 0.328]	0.049	[0.009, 0.356]	0.049	[0.020, 0.420]
9y	0.073	[0.003, 0.274]	0.080	[0.009, 0.333]	0.046	[0.011, 0.363]	0.048	[0.021, 0.430]
10y	0.068	[0.004, 0.279]	0.079	[0.009, 0.340]	0.046	[0.012, 0.370]	0.048	[0.021, 0.437]
Avg.	0.110	[0.001, 0.247]	0.112	[0.006, 0.313]	0.053	[0.005, 0.326]	0.057	[0.014, 0.400]

trend in inflation and yields. Moreover, the range of p-values obtained from our bootstrap procedure is consistent with the highly-persistent nature of trend inflation which should result in larger standard errors (Müller and Watson (2008)).

Table 4 also shows the bootstrap-based 90% confidence interval for the difference in adjusted R^2 with and without the external predictor (trend inflation). For almost all maturities and across both specifications, the confidence interval does not contain zero. Moreover, the length and distance from zero of the intervals generally agree with the magnitude of the p-values, providing a reassuring consistency of the results. This stands in contrast to the results in Bauer and Hamilton (2018, Table 2), whose bootstrap-based intervals all have a lower bound of zero regardless of the strength of evidence in favor of the external predictor based on the t-statistic.

In general, the results are supportive of the role of trend inflation as a bond risk factor, as argued by Cieslak and Povala (2015), as future returns are driven by deviations of the yield curve from a time-varying reference point determined by inflation (Rebonato and Hatano (2022)). In principle, there may be other variables that capture the trend in yields such as long-term survey forecasts of inflation expectations (Bauer and Rudebusch (2020), Crump et al. (2023)) or variables with similar trending behavior over the last 50 years. This prompts the question: is there something special about inflation (expectations)? It appears there is. As an additional exercise we considered one-sided (real-time) trend estimates based on the effective federal funds rate as a natural alternative. In unreported results, we fail to find evidence in favor of these yield-based trend estimates for bond return predictability (p-values comfortably above 10%) for the full sample.[19] This points to the possibility that inflation provides a unique (at least over this historical period) and effective real-time characterization of the low-frequency properties of bond yields up to an affine transformation.

Remark 7. There are other applications outside of trend inflation where w_t may itself represent a transformation of one or more underlying variables. For example, Cooper and Priestley (2008) utilize estimates of the output gap and Bauer and Rudebusch (2020) use estimates of the natural rate of interest. Following similar steps as above, our nonparametric bootstrap can easily accommodate such applications, including both reduced-form and structural macro-finance models. This ensures that the uncertainty arising from the generated regressors is properly accounted for in the estimated standard errors. □

4.3 Probability of Recession

The term spread – the difference between the yield on a long maturity bond and a short maturity bond – has been shown to have strong predictive power for future recessions (Harvey (1988), Estrella and Hardouvelis (1991), Chen (1991)). Historically, when the term spread is particularly compressed, this tends to be associated with a NBER-defined recession in the subsequent 2–8 quarters; of course, recessions occur relatively infrequently in the data. Against this backdrop, it is well established that limited dependent variable methods suffer from finite-sample biases when specific outcomes occur infrequently in the data (e.g., King and Zeng (2001)). We utilize our bootstrap procedure to bias-correct a probit model of future activity based on past values of the term spread.

Rather than working directly with the NBER definition of recession, we follow Rudebusch and Williams (2009) and define a recession by a real GDP contraction (negative GDP growth).[20] Let G_t be real GDP growth at time t measured at a quarterly annualized rate. To avoid any look-ahead bias, we use a time series of the third release of GDP (often referred to as the "final" release[21]) available toward the end of the subsequent quarter, rather than the current, revised series. Our z_t

[19]Full-sample detrending, which has the drawback of a look-ahead bias, performs similarly well as trend inflation measures as shown by Rebonato and Hatano (2022).

[20]This definition produces similar, but not the same, definitions of US recessions. However, along with its simplicity, it also has the advantage that this information is available toward the end of the following quarter whereas the NBER dating committee announces the designation of peaks or troughs with a considerably longer lag. See Rudebusch and Williams (2009) for additional discussion about this measure of recessions.

[21]Data are obtained from the Federal Reserve Bank of Philadelphia: `https://www.philadelphiafed.org/surveys-and-data/real-time-data-research/first-second-third`.

then becomes $z_t = \left\{ f_t^{(N)}, dr_t^{(2)}, \ldots, dr_t^{(N)}, G_t \right\}$ and we can easily block bootstrap this augmented matrix as detailed in Section 3.

The most common specification in the literature is a probit model of the form,

$$P\left(H_{t+h} = 1 | \, x_t\right) = \Phi\left(x_t' \beta\right) \tag{34}$$

for some $h \geq 1$, where H_t is a transformation of G_t. We work with two choices for H_t: first, we define $H_{1,t+h} = \mathbb{1}\left\{G_{t+h} \leq 0\right\}$; second, we define $H_{2,t+h} = \mathbb{1}\left\{\bigcup_{r=t+1}^{t+h} \left\{G_r \leq 0\right\}\right\}$. The latter measure is then set equal to one when there is a contraction in real GDP in any of the next h quarters.

Let $\hat{\beta}$ be the maximum-likelihood estimator of the coefficient β in equation (34). To construct the bootstrap-based bias correction, we block resample the matrix Z which is comprised of the stacked vectors z_t (with a pre-whitening step). In each bootstrapped sample, $(b = 1, \ldots, B)$, we first calculate $H_{t+h,b}^*$ and the requisite term structure variables, and finally $\hat{\beta}_b^*$. We can then form $\hat{d}_b^* = \hat{\beta}_b^* - \hat{\beta}$ for $b = 1, \ldots, B$. To obtain a bias-corrected estimate, we utilize

$$\hat{\beta}^{\mathrm{bc}} = \hat{\beta} - \frac{1}{B} \sum_{b=1}^{B} \hat{d}_b^*. \tag{35}$$

We can also form bootstrap-based confidence intervals by

$$C_\alpha(\beta) = \left[\hat{\beta} - \hat{d}_{(1-\alpha)}^*, \hat{\beta} - \hat{d}_{(\alpha)}^*\right], \tag{36}$$

where $\hat{d}_{(\alpha)}^*$ denotes the α-th quantile of the bootstrap distribution of $\hat{d}^*$. Note that when $\alpha = 0.5$, we obtain the bootstrap-based median unbiased estimator. For the choice of x_t, we consider two specifications. The first specification uses only the ten-year yield less the 3-month yield, which is the most common formulation used (Estrella and Hardouvelis 1991, Rudebusch and Williams 2009). The second specification then adds the 3-month yield as a separate regressor as advocated by Wright (2006). In order to be less constrained by the effective lower bound after 2008, which is not explicitly imposed in our bootstrap data, we use the deviation of the 3-month yield from its 3-year moving average.[22] This transformation also reduces the extreme persistence of the short rate and renders the dynamic properties of this predictor similar to those of the term spread. Finally, in

[22] The addition of this variable can potentially provide information on the nature ("bull" versus "bear") of the flattening or steepening of the yield curve.

order to robustify the results, we use a leave-out approach: for each t we estimate β by omitting the t-th observation and then obtain the fitted value; we follow the exact same steps in the bootstrap procedure.

Figure 9 presents the results from this exercise. We choose $h = 4$ and use $B = 999$ bootstrap samples. We use quarterly yield curve data from Gurkaynak et al. (2007) and real GDP data for the sample period 1971:Q3–2023:Q1. The top chart shows the time series of the fitted conditional probability that $H_{1t} = 1$ based on the value of the term spread. The black line depicts the estimated probability utilizing the bias-correction discussed above. We observe that the estimated probability tends to peak around the beginning of NBER recessions (grey shading). The estimate as of the beginning of 2023, is about 50% which is the highest value since the early 1980s – higher than both the 2001 and 2007-09 recessions. The second chart reflects the addition of the 3-month yield, less its 3-year moving average, as a predictor. The fitted probability is largely similar to the baseline case, although this specification is more prone to ostensibly false positives (e.g., late 1980s and 1998). Moreover, the bootstrap-based uncertainty measure around the estimated probability is wider with the extra regressor. In both specifications, the probability of recession appears elevated relative to historical levels. Finally, in the Supplemental Appendix (Section SA-2) we show the full-sample estimate versus its bias-corrected counterpart. We observe that for small fitted probabilities – periods associated with a wider term spread – the bias-corrected estimate is below the sample estimate; in contrast, for large fitted probabilities – periods associated with a compressed term spread – the bias-corrected estimate is comfortably above the sample estimate. The upward shift in the fitted probability more closely aligns with the impressive forecasting record associated with the term spread over the last 50 or so years and suggests the appropriateness of our bootstrap approach.

The bottom two charts in Figure 9, instead, show the time series of the fitted conditional probability that $H_{2t} = 1$. The third chart uses only the term spread as a predictor. The estimated probability of a contraction in real GDP occurring some time in 2023 or early 2024 is almost 90%, above the peaks in the last four cycles. The final chart shows the corresponding model with the 3-month yield deviation from its 3-year moving average as an additional predictor. The results are similar, with the upper bound of the 68% confidence interval now exceeding 90% for both specifications at the end of the sample. Finally, note that the width of the bootstrap-based confidence intervals in the top two charts is shorter than the bottom two charts. This is an appealing result of our inference approach and is consistent with the intuition that H_{2t} represents a more serially correlated process than H_{1t}, making accurate inference more challenging.

5 Conclusion

In this paper, we propose a new method for resampling the yield curve that is agnostic to the true underlying factor structure and the correct specification of the pricing model, and robust to unknown forms of serial correlation, conditional heteroskedasticity, and cross-sectional dependence. We establish the asymptotic validity of this bootstrap method under general assumptions. The primitive objects we use for resampling, excess returns and a single far-in-the-future forward rate, have more appealing statistical properties than bond prices or yields. Our approach is motivated by the fact that determining the minimal dimension of the data generating process from bond yields and returns – which is common practice in empirical work – is challenging (Crump and Gospodinov (2021)). Thus, our approach stands in sharp contrast to the conventional approach of committing to a specific parametric form for how the data were generated.

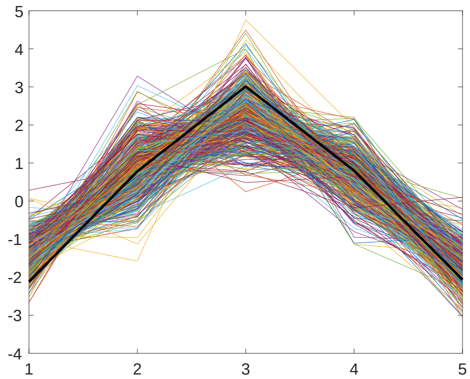
We explore the applicability of the bootstrap in the context of four empirical applications. First, we demonstrate that our nonparametric bootstrap method appears to capture accurately the "tent shape" of forward rates that was documented by Cochrane and Piazzesi (2005). Second, the model-free nature of the proposed resampling provides a unifying framework for assessing the empirical validity of the expectations hypothesis based on various regression specifications. We show how our primitive objects – obtained from deconstructing the yield curve – can be used to rewrite all of the existing specifications and propose a new regression formulation that balances the persistence properties of the dependent variable and the regressor. The empirical results across specifications offer only weak evidence against the expectations hypothesis. Third, we illustrate how to extend our bootstrap approach to bond predictive regressions with external predictors and provide support to the trend inflation factor, proposed by Cieslak and Povala (2015), in driving the low-frequency movements in bond yields. Finally, we use our bootstrap procedure to produce bias-corrected estimates and confidence intervals in probability of recession models, based on the shape of the yield curve. We observe elevated probabilities of a contraction in real GDP growth at the end of our sample.

The proposed resampling scheme can be used for generating conditional future paths of yields or other yield-related variables to obtain measures of sampling uncertainty around projected paths. This can be employed for policy analysis as well as an improved computation for option-adjusted spreads. Another advantage of our bootstrap method is in a multi-asset setup. For example, the original data matrix can be augmented with other asset returns, possibly in excess of the short rate, to ensure that the data is bootstrapped in an internally consistent manner for the purposes of predictive regressions, extracting the common factor structure of expected returns across asset classes, and other applications. These extensions are currently under investigation by the authors (Crump et al. (2021a), Crump et al. (2021b)).

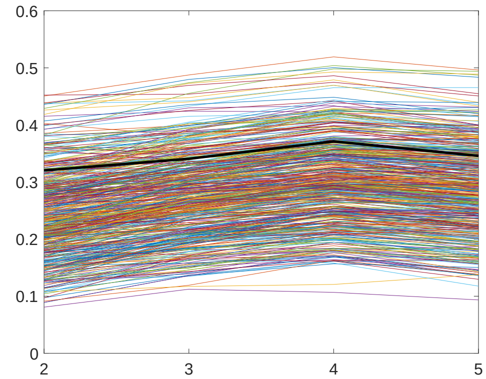
Figure 3. Resampling the Tent Shape of Cochrane and Piazzesi (2005). This figure shows the results of the nonparametric bootstrap procedure introduced in Section 3 and the parametric procedure of Bauer and Hamilton (2018). Monthly bond prices are obtained from the Fama-Bliss data set from the Center for Research in Security Prices (CRSP) and the regression specification is given in equation (10). The sample period is 1964-2003. The left column presents the sample OLS estimates (black line) and associated bootstrap estimates (multi-colored lines) for $\frac{1}{4}\sum_{n\in\{24,\cdots,60\}}(\hat{\beta}_1^{(n)},\ldots,\hat{\beta}_5^{(n)})$. The right column presents the sample R^2 (black line) and associated bootstrap estimates (multi-colored lines) across maturities.

Nonparametric Bootstrap

Regression Coefficients

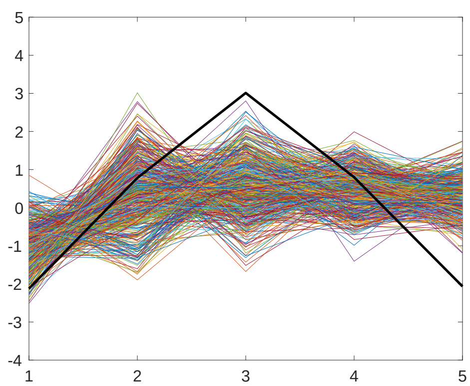


R^2

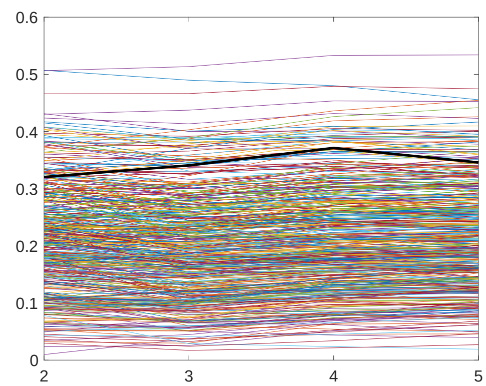


3-Factor Parametric Bootstrap

Regression Coefficients

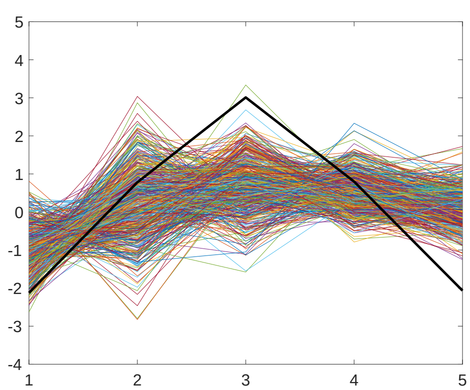


R^2



5-Factor Parametric Bootstrap

Regression Coefficients



R^2

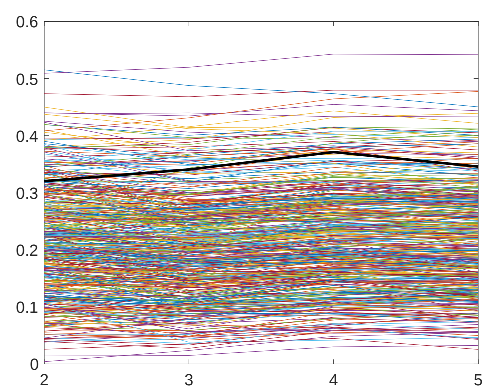


Figure 4. Realized Sample versus Bootstrapped Sample. The left column of this figure shows the time series of yields, forwards, and excess returns from Gurkaynak et al. (2007) for the sample 1972:Q1–2022:Q4. The right column of this figure shows the corresponding bootstrapped time series for these same objects.

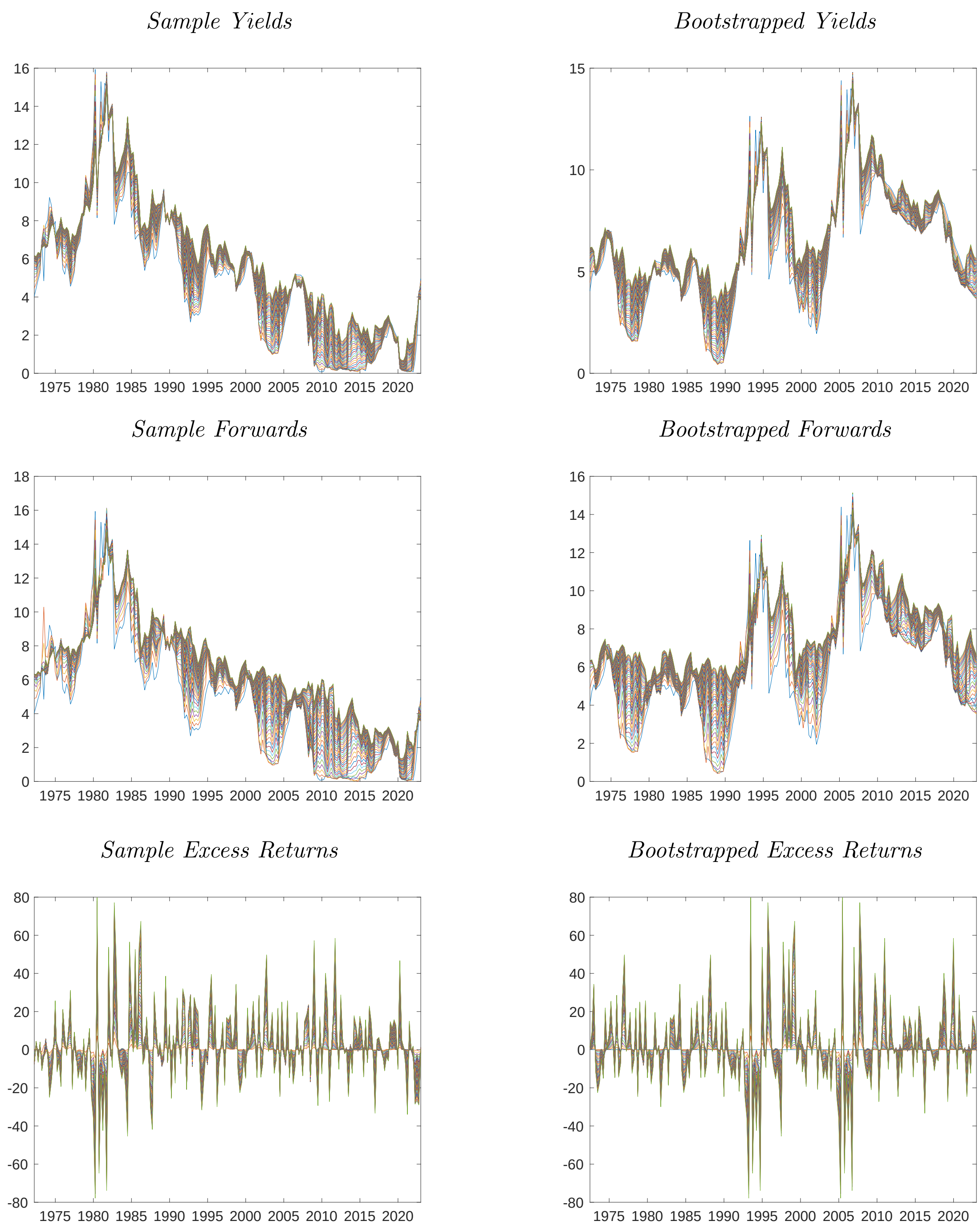


Figure 5. Expectations Hypothesis Tests: OLS Sampling Properties. This figure presents the in-sample OLS estimator and bootstrapped draws. The top row presents the time series for the CS_2 test, the middle-left plot for the CS_1 test, the middle-right plot for the FB test, and the bottom row for the CG test. The sample period is 1972:Q1–2022:Q4.

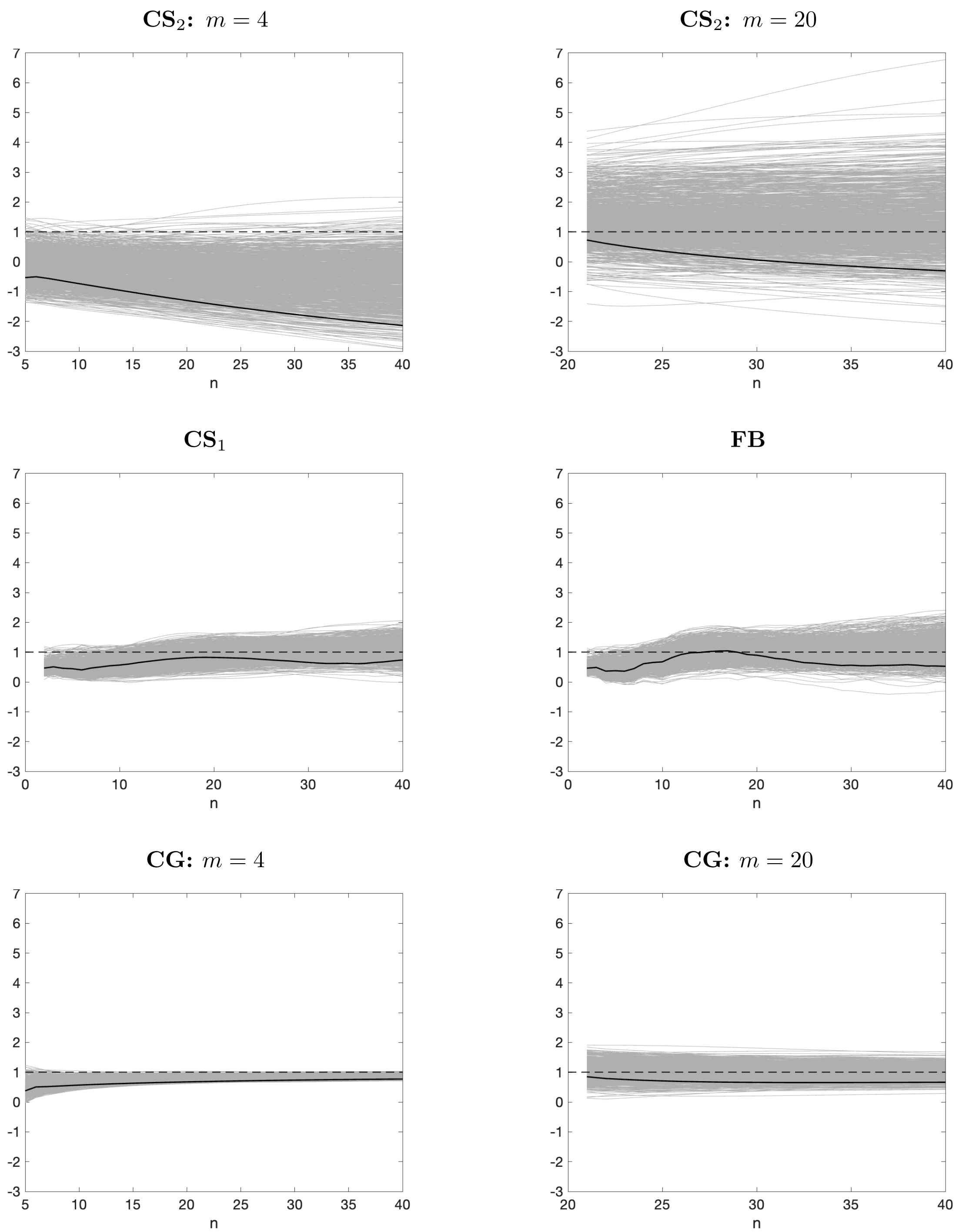


Figure 6. Expectations Hypothesis Tests: Time-Series Properties. This figure presents the dependent and explanatory variables from different regression-based tests of the expectations hypothesis. The top row presents the time series for the CS_2 test, the middle-left plot for the CS_1 test, the middle-right plot for the FB test, and the bottom row for the CG test. The sample period is 1972:Q1–2022:Q4.

Figure 7. Expectations Hypothesis Tests: Bootstrap versus Standard Inference. This figure presents p-values for the null hypothesis of $\beta = 1$ based on the nonparametric bootstrap ("Nonpar. Bootstrap") and one based on an asymptotic approximation ("Asymp. Approx."). The latter p-values are obtained using the test statistic of Lazarus et al. (2018) with the equal-weighted cosine variance estimator. Dotted horizontal lines are placed at 1%, 5% and 10% for reference. The sample period is 1972:Q1–2022:Q4.

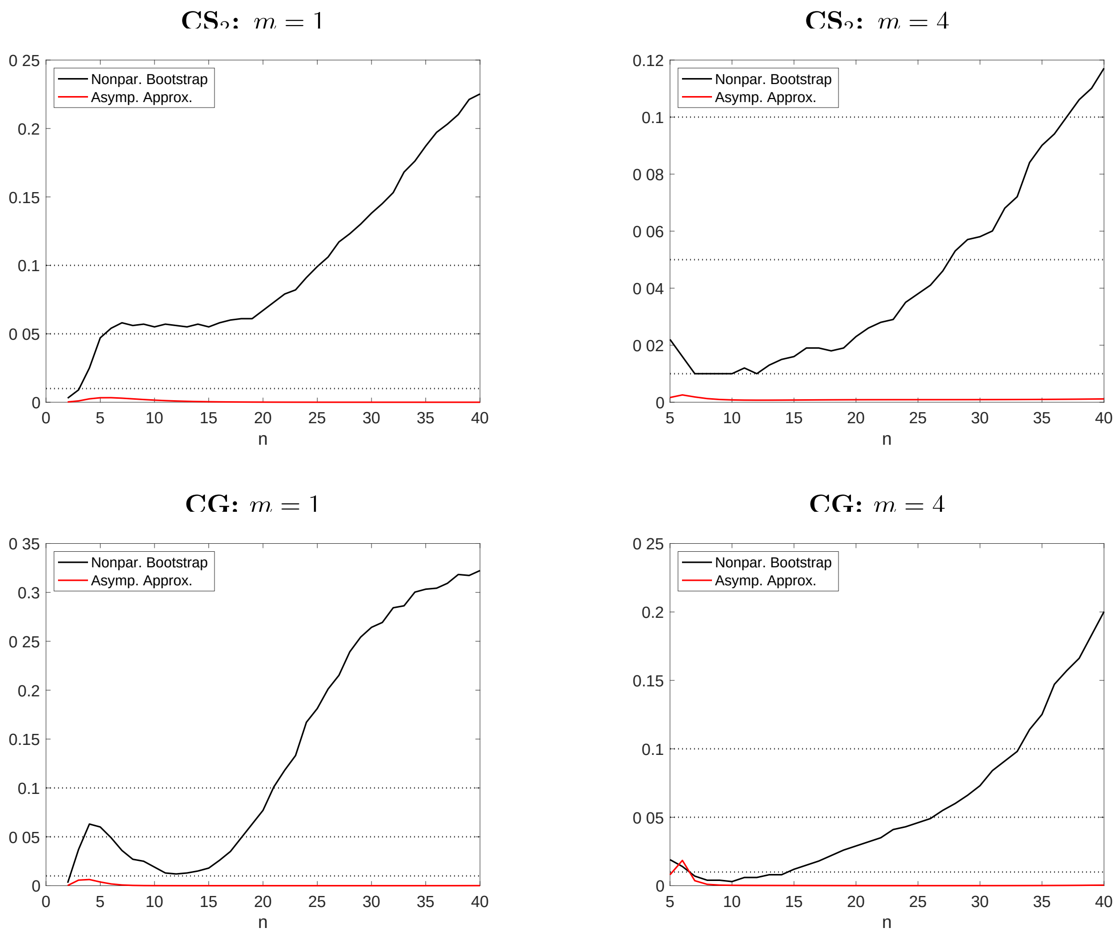


Figure 8. Realized Yields and Trend Inflation versus Bootstrapped Sample. The left chart of this figure shows the time series of yields (grey lines) and the trend inflation factor (black line) of Cieslak and Povala (2015) for the sample 1971:m11-2022:m12. The right chart of this figure shows the corresponding bootstrapped time series for these same objects.

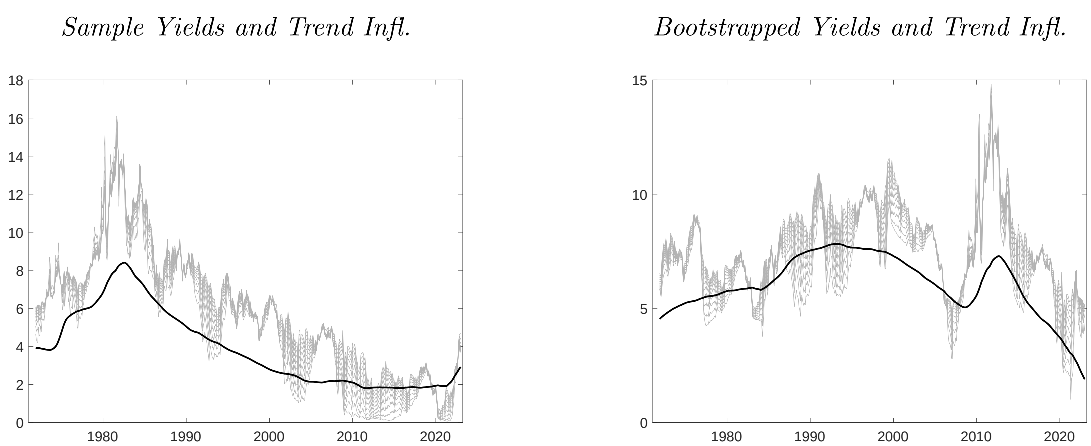
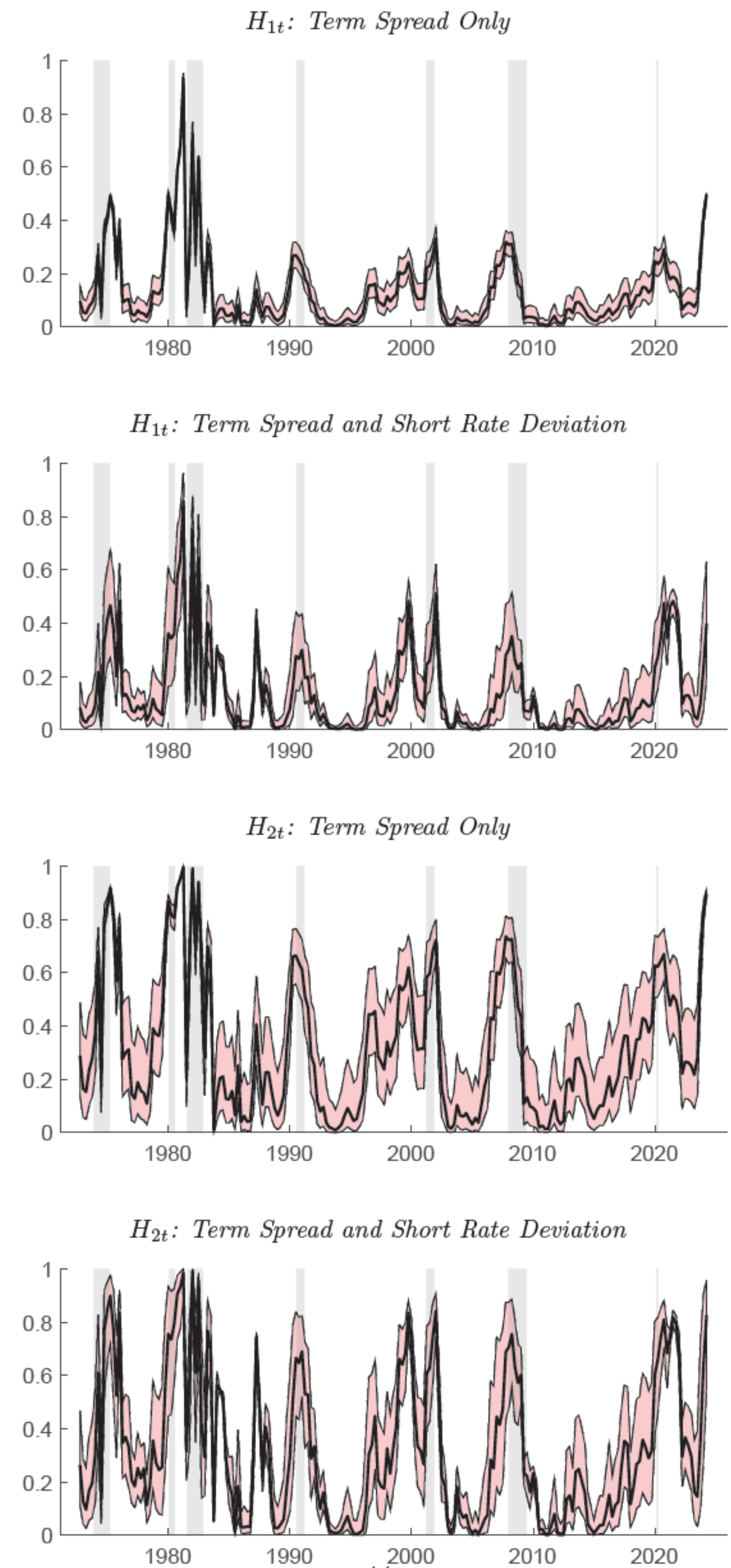


Figure 9. Probability of Recession

This figure plots the fitted values of $P(H_{1t} = 1|\, x_t)$ or $P(H_{2t} = 1|\, x_t)$ as defined in Section 4.3. x_t is either: (i) a constant and the term spread; (ii) a constant, the term spread, and the deviation of the 3-month yield from its 3-year moving average. The black line represents the estimated probability based on the bootstrap bias correction; pink shading denotes 68% pointwise confidence intervals for the associated probability. Grey shading denotes NBER recessions. The sample period is 1971:Q3–2023:Q1.



Appendix

A Assumptions and Proofs

A.1 Assumptions

Using the notation in the main text, we first provide a definition of $\{z_{nt}\}$ as a mixing process. Let $F_{-\infty}^{n,t} = \sigma(..., Z_{n,t-1}, Z_{n,t})$ and $F_{t+k}^{n,\infty} = \sigma(Z_{n,t+k}, Z_{n,t+k+1}, ...)$ denote the sigma-fields generated by the corresponding set of random variables and, for each n,

$$\alpha_n(k) = \sup_t \sup_{A \in \mathcal{F}_{-\infty}^{n,t}, B \in \mathcal{F}_{t+k}^{n,\infty}} |P(A \cap B) - P(A)P(B)|$$

Then, for each n, the random process $\{z_{nt}\}$ is α-mixing if $\alpha_n(k) \to 0$ as $k \to \infty$. Let $\|z_{nt}\|_p \equiv (E|z_{nt}|^p)^{1/p}$ denote the L_p norm of a random vector, where $|z_{nt}|$ is its Euclidean norm.

Assumption A1. Assume that

(A1.a) for each $n = 1, ..., N$, $\{z_{nt} : t = 1, ..., T\}$ are the realizations of a stationary α-mixing process with mixing coefficients $\alpha_n(k)$ such that $\sup_n \alpha_n(k) \leq \alpha(k)$, where $\alpha(k) = O(k^{-\lambda})$ for some $\lambda > (2+\delta)(r+\delta)/(r-2)$, $r > 2$ and $\epsilon > 0$;

(A1.b) for some $r > 2$ and $\epsilon > 0$, $\|z_{nt}\|_{r+\epsilon} \leq \Delta < \infty$ for all n and t;

(A1.c) $V_{NT} = \frac{1}{NT} \sum_{n=1}^{N} \sum_{t=1}^{T} E[(z_{nt} - E(z_{nt}))(z_{nt} - E(z_{nt}))']$ is positive definite uniformly in N, T;

(A1.d) N is a fixed;

(A1.e) $M_T \to \infty$ and $M_T = o(T^{1/2})$.

Assumption (A1.a) imposes restrictions on the time-series dependence without any constraints on the cross-sectional dependence. It allows for heterogeneous, but uniformly bounded, serial dependence across the different series. Stationarity along the time-series dimension can be further relaxed by allowing for some types of time heterogeneity. Assumption (A1.b) requires uniform moment bounds. While the cross-sectional (maturity) dimension N could be large relative to T and it may be convenient to allow N to be a function of T (see, for example, Valkanov (1998)), Assumption (A1.d) maintains that N is a fixed constant. Nevertheless, we use the notation $N, T \to \infty$ to accommodate the more general case when N is possibly a nondecreasing function of T. Assumption (A1.e) states that the block size M_T is allowed to grow with T but at a slower rate than $T^{1/2}$.

Assumption A2. Assume that

(A2.a) $\{(x_t', \varepsilon_{t+h}^{(n)}) : t = 1, ..., T\}$ are the realizations of a stationary α-mixing process with mixing coefficients $\alpha(k) = O(k^{-\lambda})$ for some $\lambda > 4/(r-2)$, $r > 2$;

(A2.b) for some $r > 2$ and $\epsilon > 0$, $\|x_t\|_{2(r+\epsilon)} \leq \Delta < \infty$ and $\left\|\varepsilon_{t+h}^{(n)}\right\|_{2(r+\epsilon)} \leq \Delta < \infty$ for all t;

(A2.c) $A_T = \frac{1}{T}\sum_{t=1}^{T-h}(x_t - E(x_t))(x_t - E(x_t))'$ is positive definite uniformly in T;

(A2.d) $B_T \equiv \text{Var}\left(T^{-1/2}\sum_{t=1}^{T-h}(x_t - \bar{x})\varepsilon_{t+h}^{(n)}\right)$ is $O(1)$ and $\det(B) > \epsilon$ for any $\epsilon > 0$ and $B \equiv \lim_{T\to\infty} \text{Var}\left(T^{-1/2}\sum_{t=1}^{T-h}(x_t - \bar{x})\varepsilon_{t+h}^{(n)}\right)$ is positive definite.

Assumption (A2.a) imposes stationarity on the predictors and the errors. Stationary but highly persistent predictors, such as the level factor, present challenges to statistical inference as discussed in the main text. Assumption (A2.b) provides regularity conditions that need to be strengthened further when x_t is subjected to pre-whitening by a VAR(1) model. Assumptions (A3.c) and (A3.d) are standard for the matrices A_T and B_T. Letting $\xi_{t+h} = (x_t - \bar{x})\varepsilon_{t+h}^{(n)}$, we can rewrite B_T as

$$B_T = \Gamma(0) + \sum_{j=1}^{T-h} \omega(j/b)\left(\Gamma(j) + \Gamma(j)'\right),$$

where $\Gamma(j) = E(\xi_{t+h}\xi'_{t+h-j})$, $\omega(.)$ is a kernel function and b is a bandwidth parameter. The population analog of this matrix is given by $B = \lim_{T\to\infty} \text{Var}\left(T^{-1/2}\sum_{t=1}^{T-h}\xi_{t+h}\right)$ and its HAC estimator as $\hat{B}_T = \hat{\Gamma}_T(0) + \sum_{j=1}^{T-1}\omega(j/b)(\hat{\Gamma}_T(j) + \hat{\Gamma}_T(j)')$, where $\hat{\Gamma}_T(\cdot)$ is a consistent estimator of $\Gamma(\cdot)$. The next assumption imposes sufficient conditions (see, for example, Newey and West (1987) and Andrews (1991)) for establishing consistency of the HAC estimator that is used for constructing the robust t-statistic.

Assumption A3. Assume that

(A3.a) $\omega(x)$ satisfies (i) $|\omega(x)| \leq 1$ and $\omega(x) = \omega(-x)$ for all $x \in \mathbb{R}$, (ii) $\omega(0) = 1$, (iii) $\int_{-\infty}^{\infty}|\omega(x)|\mathrm{d}x < \infty$, (iv) $\omega(x)$ is continuous at zero and almost all $x \in \mathbb{R}$, and (v) $\omega(x)$ has a characteristic exponent $k \geq 1$ which is the largest real number such that $\lim_{x\to 0}\frac{1-\omega(x)}{|x|^k} = c_k$ for some $c_k \in (0,\infty)$.

(A3.b) $b = b_T$ satisfies (i) $b_T = o(T^{1/2})$, (ii) $b_T \to \infty$ as $T \to \infty$.

Assumption (A3.a) is satisfied for popular kernels such as the Bartlett, Parzen and quadratic spectral kernels that yield positive semi-definite estimators. Assumption (A3.b) states the rate of increase for b_T that ensures the consistency of the HAC estimator (Andrews (1991)).

A.2 Proofs

Proof of Theorem 1: The limiting behavior of some terms in Theorem 1 can be inferred from the following lemma (see Goncalves (2011) for details).

Lemma A.1. *Under Assumptions A.1, A 2, and A.3,*

$$\hat{A}_T^{-1} - A_T^{-1} \xrightarrow{P} 0,$$

$$\hat{A}_T^{*-1} - \hat{A}_T^{-1} \xrightarrow{P^*} 0,$$

$$\hat{B}_T^{*} - B_T \xrightarrow{P^*} 0,$$

in probability, and

$$B_T^{-1/2}\frac{1}{\sqrt{T}}\sum_{t=1}^{T-h}(x_t^*-\bar{x}^*)\varepsilon_{t+h}^{(n)*}\xrightarrow{d^*}N(0_k,I_k),$$

where $X_T^*\xrightarrow{d^*}X$ *denotes that, conditional on the sample,* X_T^* *weakly converges to* X *under* P^*.

Next, let

$$\hat{A}_T^*=\frac{1}{T}\sum_{t=1}^{T-h}(x_t^*-\bar{x}^*)(x_t^*-\bar{x}^*)'$$

and

$$\varepsilon_{t+h}^{(n)*}=rx_{t+h}^{(n)*}-\hat{\alpha}^{(n)}-x_t^{*\prime}\hat{\beta}^{(n)}.$$

Then, we can rewrite the expression for $\sqrt{T}(\hat{\beta}^{(n)*}-\hat{\beta}^{(n)})$ as

$$\begin{aligned}\sqrt{T}(\hat{\beta}^{(n)*}-\hat{\beta}^{(n)}) &= \hat{A}_T^{*-1}\frac{1}{\sqrt{T}}\sum_{t=1}^{T-h}(x_t^*-\bar{x}^*)\varepsilon_{t+h}^{(n)*}\\ &= \left[A_T^{-1}+(\hat{A}_T^{*-1}-A_T^{-1})\right]\frac{1}{\sqrt{T}}\sum_{t=1}^{T-h}(x_t^*-\bar{x}^*)\varepsilon_{t+h}^{(n)*}\\ &= A_T^{-1}B_T^{1/2}B_T^{-1/2}\frac{1}{\sqrt{T}}\sum_{t=1}^{T-h}(x_t^*-\bar{x}^*)\varepsilon_{t+h}^{(n)*}\\ &\quad+\left[(\hat{A}_T^{*-1}-\hat{A}_T^{-1})-(\hat{A}_T^{-1}-A_T^{-1})\right]\frac{1}{\sqrt{T}}\sum_{t=1}^{T-h}(x_t^*-\bar{x}^*)\varepsilon_{t+h}^{(n)*}.\end{aligned}$$

The limiting behavior of the terms on the right-hand side is inferred from Lemma A.1. Pre-multiplying both sides by $B_T^{-1/2}A_T$ and invoking the results in Lemma A.1, we have that

$$B_T^{-1/2}A_TA_T^{-1}B_T^{1/2}B_T^{-1/2}\frac{1}{\sqrt{T}}\sum_{t=1}^{T-h}(x_t^*-\bar{x}^*)\varepsilon_{t+h}^{(n)*}\xrightarrow{d^*}N(0_k,I_k)$$

and

$$B_T^{-1/2}A_T\left[(\hat{A}_T^{*-1}-\hat{A}_T^{-1})-(\hat{A}_T^{-1}-A_T^{-1})\right]\frac{1}{\sqrt{T}}\sum_{t=1}^{T-h}(x_t^*-\bar{x}^*)\varepsilon_{t+h}^{(n)*}=o_{P^*}(1),$$

using that $\frac{1}{\sqrt{T}}\sum_{t=1}^{T-h}(x_t^*-\bar{x}^*)\varepsilon_{t+h}^{(n)*}=O_{P^*}(1)$. The first result in Theorem 1 follows from noting that, under the stated assumptions,

$$B_T^{-1/2}A_T\sqrt{T}(\hat{\beta}^{(n)}-\beta^{(n)})\xrightarrow{d}N(0_k,I_k)$$

as $T\to\infty$. The result for the t-statistic in Theorem 1 follows from Lemma A.1 and similar arguments. □

B Data Sources and Description

The primary data set we use is continuously-compounded, zero-coupon Treasury yields from Gurkaynak et al. (2007).[23] In our applications we use either monthly or quarterly data up to the ten-year maturity. In Section 2.2.1 we use the Fama-Bliss zero-coupon Treasury yields available from CRSP.[24]

Figure 1 utilizes additional data sets. First, the data on oil futures consists of WTI oil prices in USD of futures contracts (traded on NYMEX) with monthly maturities from 1 to 12 months for the period January 2000 to December 2018. The source of these data is Bloomberg. The data for the S&P500 index options is from Constantinides et al. (2013), and aggregated and sorted by maturity as in He et al. (2017).[25] The sample period is 1986:Q4 to 2012:Q1. Finally, the data on global surface temperature (measured as deviations from annual mean in Celsius) for the period 1880-2017 are obtained from the Goddard Institute for Space Studies (National Aeronautics and Space Administration) website.[26]

[23]Data are available at https://www.federalreserve.gov/pubs/feds/2006/200628/200628abs.html.
[24]The description of the data is available at https://www.crsp.org/products/documentation/fama-bliss-discount-bonds-%E2%80%93-monthly-only.
[25]We thank the authors for making the data available.
[26]https://data.giss.nasa.gov/gistemp/graphs/

References

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

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

Bauer, M. D., Rudebusch, G. D., 2020. Interest rates under falling stars. American Economic Review 110, 1316–54.

Bauer, M. D., Rudebusch, G. D., Wu, C., 2012. Correcting estimation bias in dynamic term structure models. Journal of Business and Economic Statistics 30, 454–467.

Bauer, M. D., Rudebusch, G. D., Wu, C., 2014. Term premia and inflation uncertainty: Empirical evidence from an international panel dataset: Comment,. American Economic Review 104, 323–337.

Bekaert, G., Hodrick, R. J., Marshall, D. A., 1997. On biases in tests of the expectations hypothesis of the term structure of interest rates. Journal of Financial Economics 44, 309–348.

Campbell, J. Y., 2017. Financial Decisions and Markets. Princeton University Press, Princeton, NJ.

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

Campbell, J. Y., Yogo, M., 2006. Efficient tests of stock return predictability. Journal of Financial Economics 81, 27–60.

Cavanagh, C. L., Elliott, G., Stock, J. H., 1995. Inference in models with nearly integrated regressors. Econometric Theory 11, 1131–1147.

Chen, N.-F., 1991. Financial investment opportunities and the macroeconomy. Journal of Finance 46, 529–554.

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

Cochrane, J., Piazzesi, M., 2005. Bond risk premia. American Economic Review 95, 138–160.

Cochrane, J., Piazzesi, M., 2008. Decomposing the yield curve, working paper.

Constantinides, G. M., Jackwerth, J. C., Savov, A., 2013. The puzzle of index option returns. Review of Asset Pricing Studies 3, 229–257.

Cooper, I., Priestley, R., 2008. Time-varying risk premiums and the output gap. Review of Financial Studies 22, 2801–2833.

Crump, R. K., Eusepi, S., Moench, E., Preston, B., 2023. The term structure of expectations. In: Bachmann, R., Topa, G., van der Klaauw, W. (eds.), *Handbook of Economic Expectations*, Elsevier, vol. 1, chap. 17, pp. 507–540, first ed.

Crump, R. K., Gospodinov, N., 2021. On the factor structure of bond returns. Econometrica (forthcoming).

Crump, R. K., Gospodinov, N., Tassel, P. V., 2021a. Inference on dividend and return predictability, working paper.

Crump, R. K., Gospodinov, N., Volker, D., 2021b. Inflation risk, working paper.

Davison, A. C., Hinkley, D. V., 1997. Bootstrap Methods and their Application. Cambridge University Press, Cambridge, UK.

Duffee, G. R., 2011. Information in (and not in) the term structure. Review of Financial Studies 24, 2895–2934.

Estrella, A., Hardouvelis, G. A., 1991. The term structure as a predictor of real economic activity. Journal of Finance 46, 555–576.

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

Farmer, L. E., Nakamura, E., Steinsson, J., 2022. Learning about the long run, working paper.

Ferson, W. E., Sarkissian, S., Simin, T. T., 2003. Spurious regressions in financial economics? Journal of Finance 58, 1393–1413.

Garbade, K., 1996. Fixed Income Analytics. MIT Press.

Ghysels, E., Horan, C., Moench, E., 2018. Forecasting through the rear-view mirror: Data revisions and bond return predictability. Review of Financial Studies 31, 678–714.

Giglio, S., Kelly, B., 2017. Excess volatility: Beyond discount rates. Quarterly Journal of Economics 133, 71–127.

Goncalves, S., 2011. The moving blocks bootstrap for panel linear regression models with individual fixed effects. Econometric Theory 27, 1048–1082.

Goncalves, S., Vogelsang, T. J., 2011. Block bootstrap HAC robust tests: Sophistication of the naive bootstrap. Econometric Theory 27, 745–791.

Goncalves, S., White, H., 2002. The bootstrap of the mean for dependent heterogeneous arrays. Econometric Theory 18, 1367–1384.

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

Haddad, V., Sraer, D. A., 2020. The banking view of bond risk premia. Journal of Finance 75, 2465–2502.

Harvey, C. R., 1988. The real term structure and consumption growth. Journal of Financial Economics 22, 305–333.

He, Z., Kelly, B., Manela, A., 2017. Intermediary asset pricing: New evidence from many assets. Journal of Financial Economics 126, 1–35.

Jansson, M., Moreira, M. J., 2006. Optimal inference in regression models with nearly integrated regressors. Econometrica 74, 681–714.

Joslin, S., Priebsch, M., Singleton, K. J., 2014. Risk premiums in dynamic term structure models with unspanned macro risks. Journal of Finance 69, 1197–1233.

Joslin, S., Singleton, K. J., Zhu, H., 2011. A new perspective on Gaussian dynamic term structure models. Review of Financial Studies 24, 926–970.

Kiefer, N. M., Vogelsang, T. J., 2005. A new asymptotic theory for heteroskedasticity-autocorrelation robust tests. Econometric Theory 21, 1130–1164.

Kilian, L., 1998. Small-sample confidence intervals for impulse response functions. Review of Economics and Statistics 80, 218–230.

King, G., Zeng, L., 2001. Logistic regression in rare events data. Political Analysis 9, 137–163.

Lazarus, E., Lewis, D. J., Stock, J. H., Watson, M. W., 2018. HAR inference: Recommendations for practice. Journal of Business & Economic Statistics 36, 541–559.

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

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

McCarthy, D., Zhang, K., Brown, L., Berk, R., Buja, A., George, E., Zhao, L., 2018. Calibrated percentile double bootstrap for robust linear regression inference. Statistica Sinica 4, 2565–2589.

Müller, U. K., Watson, M. W., 2008. Testing models of low-frequency variability. Econometrica 76, 979–1016.

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

Niebuhr, T., Kreiss, J.-P., Paparoditis, E., 2017. Some properties of the autoregressive-aided block bootstrap. Electronic Journal of Statistics 11, 725–751.

Olea, J. L. M., Plagborg-Møller, M., 2021. Local projection inference is simpler and more robust than you think. Econometrica 89, 1789–1823.

Onatski, A., Wang, C., 2021. Spurious factor analysis. Econometrica 89, 591–614.

Politis, D. N., Romano, J. P., 1994. The stationary bootstrap. Journal of the American Statistical Association 89, 1303–1313.

Rebonato, R., Hatano, T., 2022. Why does the Cieslak-Povala model predict treasury returns? a reinterpretation. Journal of Fixed Income 33, 20–32.

Rebonato, R., Zanetti, P., 2023. Does the Cochrane-Piazzesi factor predict? an international resampling perspective. Journal of Fixed Income 32, 1–16.

Rossi, B., 2007. Expectations hypotheses tests at long horizons. Econometrics Journal 10, 554–579.

Rudebusch, G. D., Williams, J. C., 2009. Forecasting recessions: The puzzle of the enduring power of the yield curve. Journal of Business & Economic Statistics 27, 492–503.

Speck, C., 2021. Pricing the bund term structure with linear regressions - without an observable short rate, working paper.

Stambaugh, R. F., 1999. Predictive regressions. Journal of Financial Economics 54, 375–421.

Stock, J. H., Watson, M. W., 2007. Why has u.s. inflation become harder to forecast? Journal of Money, Credit and Banking 39, 3–33.

Uhlig, H., 2009. Comment on 'how has the Euro changed the monetary transmission mechanism?'. In: Acemoglu, D., Rogoff, K., Woodford, M. (eds.), *NBER Macroeconomics Annual*, MIT Press, vol. 24, pp. 141–152.

Valkanov, R., 1998. The term structure with highly persistent interest rates, working paper.

Wei, M., Wright, J. H., 2013. Reverse regressions and long-horizon forecasting. Journal of Applied Econometrics 28, 353–371.

Wright, J. H., 2006. The yield curve and predicting recessions. Finance and Economics Discussion Series 2006-7, Federal Reserve Board.

Supplement to "Deconstructing the Yield Curve"

August 23, 2023

SA-1 Additional Simulation Results

This appendix provides additional simulation results exploring the properties of our bootstrap procedure and comparing to alternatives in the literature.

Tables SA.1, SA.2, SA.4, SA.6, and SA.7 present results from HAR approaches based on asymptotic approximations of the corresponding test statistics. We include results using the variance estimator of Newey and West (1987) with lag length h (labelled "NW") along with the two inference procedures discussed in Lazarus et al. (2018). The first approach uses the variance estimator of Newey and West (1987) and fixed-b asymptotic approximation (labelled "LLSW-NW"). The second approach uses the equal-weighted cosine (EWC) estimator of the long-run variance and a limiting t distribution (labelled "LLSW-EWC"). We implement both of these tests exactly as described in Lazarus et al. (2018). We do note that Lazarus et al. (2018, p. 542) do not recommend the use of these tests for high degrees of persistence as in our case. Instead, we include these results as a benchmark comparison only.

The first two tables in this appendix, Tables SA.1 and SA.2, present the NW, LLSW-NW, and LLSW-EWC results for the designs presented in Tables 1 and 2 in the main text, respectively. Table SA.7 presents the NW, LLSW-NW, and LLSW-EWC results for the design presented in Table 3 in the main text.

Tables SA.3 and SA.4 present the results for the nonparametric bootstrap, the parametric bootstrap of Bauer and Hamilton (2018), NW, LLSW-NW, and LLSW-EWC for the case of three forward factors which follow a VAR(2) as discussed in Section 2.2.

Tables SA.5 and Table SA.6 present the results for the nonparametric bootstrap, the parametric bootstrap of Bauer and Hamilton (2018), NW, LLSW-NW, and LLSW-EWC for the case of three yield factors which follow a VAR(1) as discussed in Section 2.2.

Table SA.1. Simulations (VAR(1) Specification, 3 Forward Factors) This table presents empirical size for the methods based on asymptotic approximations as described in the main text. The nominal level is 10% and the sample size is $T = 600$. Each column reports results for the t-test associated with the regressor $(g_{1t}, g_{2t}, g_{3t}) = (y_t^{(12)}, f_t^{(60,12)}, f_t^{(120,12)})$. Based on 1,000 simulations and 399 bootstrap replications per simulation.

	NW			**LLSW-NW**			**LLSW-EWC**		
		Size			*Size*			*Size*	
Maturity	g_{1t}	g_{2t}	g_{3t}	g_{1t}	g_{2t}	g_{3t}	g_{1t}	g_{2t}	g_{3t}
2y	0.244	0.197	0.206	0.182	0.135	0.143	0.197	0.147	0.155
3y	0.254	0.194	0.212	0.193	0.143	0.147	0.204	0.156	0.158
4y	0.247	0.201	0.202	0.179	0.145	0.150	0.194	0.158	0.162
5y	0.245	0.204	0.220	0.179	0.152	0.150	0.189	0.164	0.165
6y	0.238	0.208	0.214	0.171	0.147	0.157	0.182	0.161	0.168
7y	0.240	0.211	0.217	0.174	0.143	0.157	0.182	0.161	0.163
8y	0.238	0.210	0.225	0.174	0.139	0.158	0.185	0.159	0.166
9y	0.239	0.203	0.214	0.171	0.144	0.158	0.182	0.162	0.170
10y	0.238	0.200	0.210	0.172	0.142	0.158	0.181	0.160	0.163

Table SA.2. Simulations (VAR(1) Specification, 5 Forward Factors) This table presents empirical size and power for the methods based on asymptotic approximations as described in the main text. The nominal level is 10% and the sample size is $T = 600$. Each column reports results for the t-test associated with the regressor $(g_{1t}, g_{2t}, g_{3t}, g_{4t}, g_{5t}) = (y_t^{(12)}, f_t^{(36,12)}, f_t^{(60,12)}, f_t^{(84,12)}, f_t^{(120,12)})$. Based on 1,000 simulations and 399 bootstrap replications per simulation.

	NW						**LLSW-NW**				
	Size						*Size*				
Maturity	g_{1t}	g_{2t}	g_{3t}	g_{4t}	g_{5t}	Maturity	g_{1t}	g_{2t}	g_{3t}	g_{4t}	g_{5t}
2y	0.209	0.169	0.167	0.172	0.187	2y	0.160	0.146	0.136	0.143	0.152
3y	0.195	0.170	0.163	0.171	0.179	3y	0.153	0.136	0.134	0.150	0.156
4y	0.183	0.175	0.169	0.168	0.185	4y	0.146	0.141	0.144	0.148	0.159
5y	0.187	0.176	0.176	0.180	0.195	5y	0.159	0.139	0.161	0.147	0.163
6y	0.201	0.184	0.187	0.177	0.198	6y	0.160	0.148	0.163	0.149	0.164
7y	0.206	0.184	0.188	0.183	0.198	7y	0.171	0.143	0.162	0.153	0.162
8y	0.216	0.190	0.193	0.191	0.199	8y	0.168	0.147	0.161	0.166	0.164
9y	0.214	0.196	0.198	0.194	0.198	9y	0.170	0.152	0.163	0.167	0.161
10y	0.208	0.194	0.191	0.193	0.192	10y	0.173	0.156	0.165	0.167	0.161

	LLSW-EWC				
	Size				
Maturity	g_{1t}	g_{2t}	g_{3t}	g_{4t}	g_{5t}
2y	0.155	0.137	0.130	0.139	0.141
3y	0.140	0.131	0.121	0.130	0.143
4y	0.138	0.138	0.141	0.131	0.150
5y	0.143	0.133	0.147	0.138	0.158
6y	0.148	0.138	0.149	0.145	0.160
7y	0.153	0.138	0.158	0.149	0.157
8y	0.152	0.138	0.156	0.151	0.153
9y	0.157	0.147	0.158	0.153	0.153
10y	0.163	0.148	0.159	0.158	0.152

Table SA.3. Simulations (VAR(2) Specification, 3 Forward Factors) This table presents empirical size and power for the nonparametric and parametric bootstrap methods described in the main text. The nominal level is 10% and the sample size is $T = 600$. Each column reports results for the t-test associated with the regressor $(g_{1t}, g_{2t}, g_{3t}, g_{4t}, g_{5t}, g_{6t}) = (y_t^{(12)}, f_t^{(60,12)}, f_t^{(120,12)}, y_{t-1}^{(12)}, f_{t-1}^{(60,12)}, f_{t-1}^{(120,12)})$. Based on 1,000 simulations and 399 bootstrap replications per simulation.

Nonparametric Bootstrap

	Size						*Power*					
Maturity	g_{1t}	g_{2t}	g_{3t}	g_{4t}	g_{5t}	g_{6t}	g_{1t}	g_{2t}	g_{3t}	g_{4t}	g_{5t}	g_{6t}
2y	0.119	0.132	0.101	0.097	0.076	0.073	0.228	0.226	0.100	0.119	0.080	0.074
3y	0.117	0.133	0.112	0.098	0.084	0.073	0.378	0.324	0.123	0.114	0.089	0.070
4y	0.120	0.130	0.118	0.107	0.085	0.075	0.474	0.399	0.141	0.108	0.092	0.071
5y	0.115	0.132	0.121	0.109	0.089	0.071	0.534	0.415	0.130	0.103	0.092	0.072
6y	0.115	0.126	0.122	0.111	0.090	0.072	0.577	0.389	0.124	0.103	0.090	0.073
7y	0.114	0.127	0.122	0.110	0.091	0.072	0.605	0.346	0.129	0.102	0.092	0.074
8y	0.114	0.127	0.129	0.102	0.088	0.072	0.612	0.306	0.147	0.098	0.094	0.074
9y	0.113	0.123	0.126	0.104	0.092	0.072	0.614	0.245	0.173	0.095	0.097	0.072
10y	0.108	0.118	0.125	0.106	0.095	0.074	0.606	0.195	0.226	0.092	0.100	0.073

Parametric Bootstrap

	Size						*Power*					
	g_{1t}	g_{2t}	g_{3t}	g_{4t}	g_{5t}	g_{6t}	g_{1t}	g_{2t}	g_{3t}	g_{4t}	g_{5t}	g_{6t}
2y	0.039	0.054	0.011	0.026	0.000	0.000	0.063	0.138	0.007	0.000	0.000	0.000
3y	0.033	0.055	0.009	0.030	0.000	0.000	0.182	0.196	0.025	0.000	0.000	0.000
4y	0.036	0.041	0.001	0.030	0.000	0.000	0.266	0.259	0.023	0.000	0.000	0.000
5y	0.031	0.049	0.001	0.021	0.000	0.000	0.359	0.267	0.007	0.000	0.000	0.000
6y	0.033	0.031	0.000	0.021	0.000	0.000	0.381	0.303	0.000	0.000	0.000	0.000
7y	0.032	0.023	0.001	0.020	0.000	0.000	0.399	0.276	0.000	0.000	0.000	0.000
8y	0.033	0.019	0.000	0.019	0.000	0.000	0.411	0.217	0.000	0.000	0.000	0.000
9y	0.032	0.023	0.001	0.020	0.000	0.000	0.411	0.182	0.000	0.000	0.000	0.000
10y	0.032	0.011	0.000	0.015	0.000	0.000	0.402	0.096	0.000	0.000	0.000	0.000

Table SA.4. Simulations (VAR(2) Specification, 3 Forward Factors) This table presents empirical size and power for the methods based on asymptotic approximations as described in the main text. The nominal level is 10% and the sample size is $T = 600$. Each column reports results for the t-test associated with the regressor $(g_{1t}, g_{2t}, g_{3t}, g_{4t}, g_{5t}, g_{6t}) = (y_t^{(12)}, f_t^{(60,12)}, f_t^{(120,12)}, y_{t-1}^{(12)}, f_{t-1}^{(60,12)}, f_{t-1}^{(120,12)})$. Based on 1,000 simulations and 399 bootstrap replications per simulation.

NW

Size

Maturity	g_{1t}	g_{2t}	g_{3t}	g_{4t}	g_{5t}	g_{6t}
2y	0.160	0.168	0.167	0.134	0.124	0.152
3y	0.156	0.168	0.167	0.131	0.136	0.152
4y	0.150	0.169	0.177	0.138	0.141	0.146
5y	0.146	0.166	0.186	0.141	0.139	0.148
6y	0.147	0.163	0.180	0.141	0.138	0.147
7y	0.147	0.168	0.179	0.147	0.139	0.151
8y	0.140	0.161	0.176	0.150	0.137	0.152
9y	0.140	0.159	0.169	0.145	0.138	0.153
10y	0.135	0.156	0.168	0.145	0.142	0.157

LLSW-NW

Size

Maturity	g_{1t}	g_{2t}	g_{3t}	g_{4t}	g_{5t}	g_{6t}
2y	0.155	0.150	0.138	0.133	0.116	0.131
3y	0.151	0.152	0.146	0.140	0.118	0.129
4y	0.144	0.151	0.153	0.142	0.123	0.131
5y	0.141	0.147	0.156	0.136	0.118	0.137
6y	0.136	0.146	0.154	0.135	0.121	0.133
7y	0.135	0.143	0.159	0.137	0.130	0.135
8y	0.129	0.142	0.155	0.141	0.129	0.134
9y	0.126	0.139	0.153	0.141	0.131	0.134
10y	0.129	0.140	0.152	0.143	0.127	0.137

LLSW-EWC

Size

Maturity	g_{1t}	g_{2t}	g_{3t}	g_{4t}	g_{5t}	g_{6t}
2y	0.151	0.145	0.137	0.133	0.107	0.124
3y	0.145	0.147	0.149	0.131	0.109	0.129
4y	0.140	0.150	0.157	0.131	0.118	0.133
5y	0.140	0.146	0.159	0.134	0.114	0.133
6y	0.134	0.144	0.157	0.137	0.113	0.130
7y	0.138	0.141	0.157	0.136	0.120	0.130
8y	0.130	0.137	0.153	0.136	0.121	0.130
9y	0.127	0.136	0.151	0.140	0.123	0.129
10y	0.126	0.138	0.151	0.138	0.122	0.136

Table SA.5. Simulations (VAR(1) Specification, 3 Yield Factors) This table presents empirical size and power for the nonparametric and parametric bootstrap methods described in the main text. The nominal level is 10% and the sample size is $T = 600$. Each column reports results for the t-test associated with the regressor $(g_{1t}, g_{2t}, g_{3t}) = (y_t^{(12)}, y_t^{(60)}, y_t^{(120)})$. Based on 1,000 simulations and 399 bootstrap replications per simulation.

	Nonparametric Bootstrap					
	Size			*Power*		
Maturity	g_{1t}	g_{2t}	g_{3t}	g_{1t}	g_{2t}	g_{3t}
2y	0.095	0.069	0.067	0.133	0.080	0.192
3y	0.105	0.083	0.075	0.198	0.071	0.169
4y	0.099	0.087	0.075	0.259	0.079	0.171
5y	0.095	0.092	0.074	0.273	0.074	0.203
6y	0.096	0.088	0.077	0.264	0.077	0.260
7y	0.098	0.085	0.088	0.250	0.083	0.327
8y	0.098	0.087	0.085	0.226	0.113	0.398
9y	0.098	0.089	0.093	0.206	0.148	0.459
10y	0.095	0.089	0.093	0.187	0.179	0.519
	Parametric Bootstrap					
	Size			*Power*		
Maturity	g_{1t}	g_{2t}	g_{3t}	g_{1t}	g_{2t}	g_{3t}
2y	0.023	0.032	0.038	0.080	0.025	0.121
3y	0.023	0.035	0.036	0.144	0.026	0.099
4y	0.021	0.034	0.034	0.196	0.023	0.119
5y	0.021	0.029	0.030	0.215	0.026	0.163
6y	0.021	0.029	0.031	0.200	0.022	0.233
7y	0.020	0.026	0.033	0.182	0.039	0.329
8y	0.019	0.023	0.038	0.145	0.069	0.419
9y	0.020	0.022	0.043	0.116	0.107	0.487
10y	0.015	0.022	0.051	0.095	0.158	0.540

Table SA.6. Simulations (VAR(1) Specification, 3 Yield Factors) This table presents empirical size and power for the methods based on asymptotic approximations as described in the main text. The nominal level is 10% and the sample size is $T = 600$. Each column reports results for the t-test associated with the regressor $(g_{1t}, g_{2t}, g_{3t}) = (y_t^{(12)}, y_t^{(60)}, y_t^{(120)})$. Based on 1,000 simulations and 399 bootstrap replications per simulation.

	NW				**LLSW-NW**				**LLSW-EWC**		
	Size				*Size*				*Size*		
Maturity	g_{1t}	g_{2t}	g_{3t}	Maturity	g_{1t}	g_{2t}	g_{3t}	Maturity	g_{1t}	g_{2t}	g_{3t}
2y	0.214	0.193	0.212	2y	0.163	0.154	0.169	2y	0.137	0.141	0.159
3y	0.207	0.210	0.210	3y	0.154	0.163	0.172	3y	0.131	0.149	0.149
4y	0.204	0.224	0.221	4y	0.152	0.161	0.172	4y	0.134	0.149	0.160
5y	0.202	0.219	0.227	5y	0.158	0.164	0.174	5y	0.140	0.153	0.166
6y	0.198	0.221	0.227	6y	0.162	0.163	0.181	6y	0.144	0.149	0.166
7y	0.202	0.218	0.229	7y	0.160	0.163	0.174	7y	0.144	0.145	0.164
8y	0.203	0.216	0.230	8y	0.158	0.156	0.175	8y	0.146	0.143	0.162
9y	0.199	0.201	0.227	9y	0.159	0.148	0.181	9y	0.145	0.135	0.160
10y	0.201	0.197	0.226	10y	0.161	0.150	0.183	10y	0.144	0.136	0.162

Table SA.7. Simulations (VAR(1) Specification, 3 Yield Factors, 2 Macro Predictors) This table presents empirical size and power for the methods based on asymptotic approximations as described in the main text. The nominal level is 10% and the sample size is $T = 600$. Each column reports results for the t-test associated with the regressor $(g_{1t}, g_{2t}, g_{3t}) = (y_t^{(12)}, y_t^{(60)}, y_t^{(120)})$. For the bivariate external regressors, w_t, only empirical size is reported. Based on 1,000 simulations and 399 bootstrap replications per simulation.

	NW						**LLSW-NW**				
	Size						*Size*				
Maturity	g_{1t}	g_{2t}	g_{3t}	w_{1t}	w_{2t}	Maturity	g_{1t}	g_{2t}	g_{3t}	w_{1t}	w_{2t}
2y	0.224	0.178	0.194	0.230	0.170	2y	0.172	0.147	0.150	0.184	0.130
3y	0.224	0.184	0.202	0.233	0.166	3y	0.168	0.147	0.150	0.184	0.134
4y	0.209	0.191	0.200	0.233	0.168	4y	0.162	0.154	0.150	0.182	0.142
5y	0.218	0.196	0.198	0.240	0.170	5y	0.162	0.163	0.153	0.186	0.140
6y	0.216	0.197	0.196	0.238	0.168	6y	0.161	0.159	0.151	0.193	0.139
7y	0.214	0.198	0.195	0.250	0.166	7y	0.169	0.164	0.161	0.197	0.135
8y	0.221	0.194	0.205	0.250	0.168	8y	0.168	0.165	0.165	0.196	0.146
9y	0.219	0.197	0.207	0.245	0.170	9y	0.167	0.162	0.166	0.200	0.147
10y	0.213	0.199	0.215	0.244	0.171	10y	0.173	0.165	0.168	0.197	0.146

	LLSW-EWC				
	Size				
Maturity	g_{1t}	g_{2t}	g_{3t}	w_{1t}	w_{2t}
2y	0.157	0.134	0.146	0.174	0.131
3y	0.159	0.143	0.140	0.173	0.122
4y	0.160	0.142	0.139	0.173	0.128
5y	0.153	0.148	0.138	0.181	0.131
6y	0.152	0.148	0.137	0.177	0.129
7y	0.159	0.147	0.149	0.182	0.133
8y	0.152	0.147	0.150	0.186	0.135
9y	0.152	0.145	0.153	0.188	0.140
10y	0.154	0.147	0.157	0.192	0.138

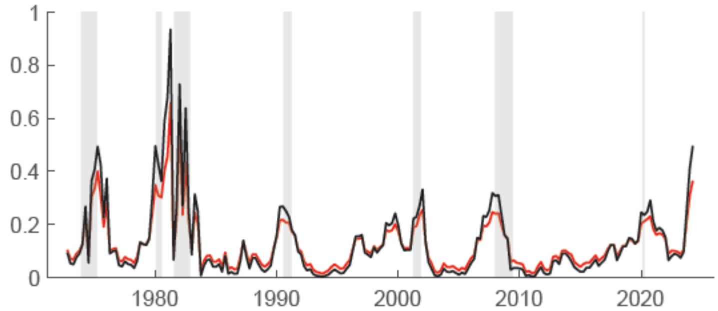
SA-2 Probability of Recession Based on the Term Spread

Figure SA.1 compares the full-sample (leave-one-out) standard probit estimator (red line) along with the bootstrap-bias corrected (leave-one-out) probit estimator (black line). Across both specifications and choices of horizon we observe that the bias-corrected estimate is lower when the probability of recession is estimated to be low but it is higher when the probability of recession is estimated to be high. This results in more variation of the bias-corrected estimate with much more prominent peaks in the estimated probability of recession before NBER recessions. As noted in the main text, this upward shift in the fitted probabilities more closely aligns with the impressive forecasting record associated with the term spread over the last 50 or so years and suggests the appropriateness of our bootstrap approach.

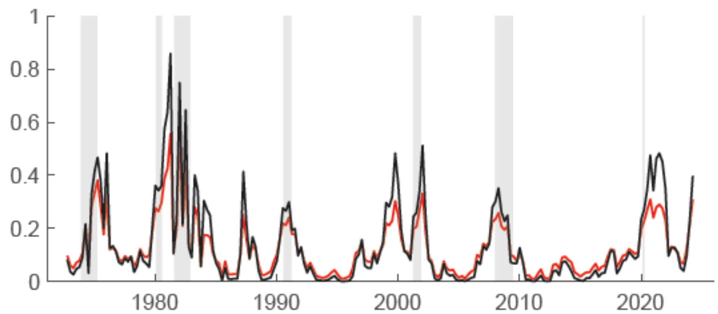
Figure SA.1. Probability of Recession

This figure plots the fitted values of $P(H_{1t} = 1|\, x_t)$ or $P(H_{2t} = 1|\, x_t)$ as defined in the main text. x_t is either: (i) a constant and the term spread; (ii) a constant, the term spread, and the deviation of the 3-month yield from its 3-year moving average. The black line represents the estimated probability based on the bootstrap bias correction and the red line represents the full sample estimate. Grey shading denotes NBER recessions. The sample period is 1971:Q3–2023:Q1.

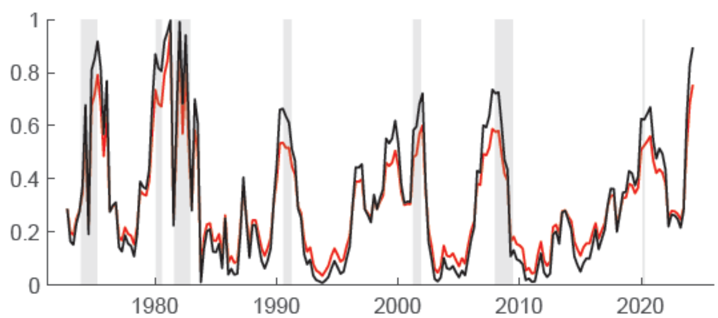
H_{1t}: Term Spread Only



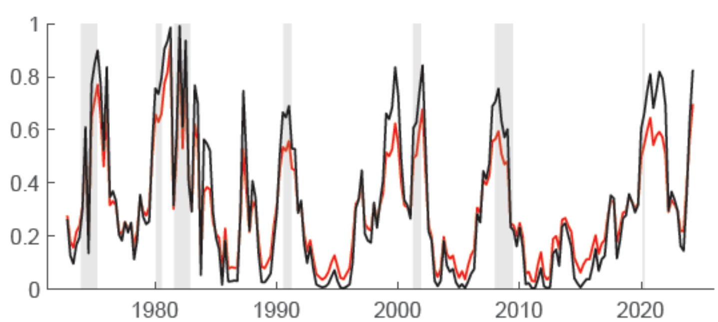
H_{1t}: Term Spread and Short Rate Deviation



H_{2t}: Term Spread Only



H_{2t}: Term Spread and Short Rate Deviation



References

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

Lazarus, E., Lewis, D. J., Stock, J. H., Watson, M. W., 2018. HAR inference: Recommendations for practice. Journal of Business & Economic Statistics 36, 541–559.

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