STAFF
REPORTS

NO. 884
APRIL 2019

REVISED
AUGUST 2024

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 2024
JEL classification: G10, G12, C15, C58

Abstract

We introduce a novel nonparametric bootstrap for the yield curve which is agnostic to the true factor structure of interest rates. We deconstruct the yield curve into primitive objects, with weak cross-sectional and time-series dependence, that serve as building blocks for resampling the data. We analyze the properties of the bootstrap for mimicking salient features of the data and conducting valid inference. We demonstrate the benefits of our general method by revisiting the predictability of bond returns based on slow-moving fundamentals. We find that trend inflation, but not the equilibrium real rate, has predictive power for future bond returns.

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), Keshav Dogra, 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.

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 lead to spurious commonality across series suggestive of a stronger factor structure than might exist. Crump and Gospodinov (2022) 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 inherently distinct from the usual linear regression setup and implies that standard 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 model-free 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. Importantly, our results translate directly to all other assets with a finite maturity structure such as options, swaps, or futures.

A key feature of our method is that the term structure of interest rates can be fully reconstructed from the bootstrapped primitive objects in an internally consistent manner via a set of identities. As the resampled data naturally satisfies term structure identities, this bottom-up approach ensures that any predictability in future returns from past yields or forwards is retained. This property is particularly advantageous when 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; Bauer and Rudebusch 2020; Bauer and Chernov 2024). One prominent example is the equilibrium real rate, or r^*, which is an explicit function of short-term interest rates and must be resampled in such a way that is consistent with the rest of the nominal yield curve. Our method is also well-suited to accommodate 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) by preserving their potential relationships with the yield curve and accounting for sampling uncertainty, arising from prior data transformations such as exponential smoothing.

We provide extensive Monte Carlo simulation evidence illustrating the ability of our method

to retain the unknown factor structure in the data. The simulation experiments also highlight the extremely challenging nature of conducting inference in bond predictive regressions that is not fully appreciated in the literature. We show that our bootstrap inference procedure controls size well for multi-period holding returns at various maturities and across different specifications. Additionally, the simulation results suggest that this size control is achieved with only a minimal loss of power relative to an infeasible "oracle" method which is based on the exact structure and the true parameters of the data generating process.

In our main empirical application, we revisit the results from Cieslak and Povala (2015) and Bauer and Rudebusch (2020) regarding the predictive content in trend inflation and the equilibrium real interest rate (r^*) for future bond returns. Using our bootstrap, we find compelling evidence that trend inflation has additional explanatory power for future bond returns beyond what is captured by the current yield curve. In contrast, we fail to find any evidence that r^* has predictive power; importantly, this result holds uniformly across different subsamples, maturities, and regression specifications. Furthermore, we also clarify why the bootstrap method of Bauer and Hamilton (2018), which Bauer and Rudebusch (2020) rely on, is invalid in this setting.

We consider two additional empirical applications which highlight the broad applicability of our bootstrap method. First, we revisit the regression-based tests of the expectations hypothesis (Fama and Bliss 1987, Campbell and Shiller 1991).[1] This is a perfect setting to demonstrate the utility of our bootstrap as the dependent and independent variables are both functions of yields. Furthermore, since we jointly bootstrap the whole yield curve, we can assess all of the different regression-based tests of the expectations hypothesis in a unified way. We show how our bootstrap enhances our understanding of the sampling properties of the OLS estimator in this setting. Second, we use our new bootstrap procedure to construct bias-corrected estimates and confidence intervals for the probability of recession, using information embedded in the yield curve. We show that 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.

The rest of the paper is organized as follows. In Section 2, we introduce the nonparametric bootstrap method. We demonstrate the appealing finite-sample properties of our bootstrap in a series of simulation experiments in Section 3. Section 4 presents results from our three empirical

[1] Of independent interest, we also show how to view all of the existing regression specifications for testing the expectations hypothesis in terms of a single unifying primitive equation.

applications. Section 5 concludes. Additional details and results are available in Appendix A and a Supplemental Appendix (hereafter, SA).

2 Resampling the Yield Curve

2.1 The Need for Robust Inference: A Motivation

We start this section by introducing notation used throughout the paper. We also provide a heuristic motivation for why inference methods that exhibit robustness to the unknown factor structure and time-series properties of bond data are desirable.

Let $p_t^{(n)}$ denote 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 $p_t^{(0)} = 0$, 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

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

It is well known that bond yields have both strong cross-sectional and time-series dependence which makes inference and, in particular, resampling-based inference challenging. We will begin by discussing how the special setting of the yield curve is reflected in the strong cross-sectional dependence. First, observe that, by equation (2), $p_t^{(n)}$ and $y_t^{(n)}$ are cross-sectional partial sums and partial averages of forward rates, respectively. These formulas demonstrate that two bond prices or yields, $y_t^{(n)}$ and $y_t^{(m)}$, have $\min(m, n)$ forwards in common. 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 exhibit stronger local correlation than forwards across the same maturities.[2]

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

What are the implications of these cross-sectional restrictions for characterizing the true factor space? It is now ubiquitous to estimate the factor structure of the term structure by applying static principal component (PC) 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.[3] 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.

A similar pattern appears to hold for any asset with a finite maturity structure. In the top right and bottom left panels we show the associated PC loadings for oil futures and S&P 500 options returns, respectively. However, in the last panel we present the PC loadings from a panel data set of the seasonal cycle of global surface temperature. Remarkably again, the loadings are very similar in shape to those estimated from maturity-ordered financial assets. This provides 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. In fact, Crump and Gospodinov (2022) show that this behavior can be consistent with a much higher dimensional factor space.

But cross-sectional dependence is not the only challenge. Both bond yields and forward rates exhibit a high degree of time-series persistence. 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 which is naturally achieved in this setting through bond returns. Define the one-period holding return on a bond of maturity n from time t to $t+1$ as

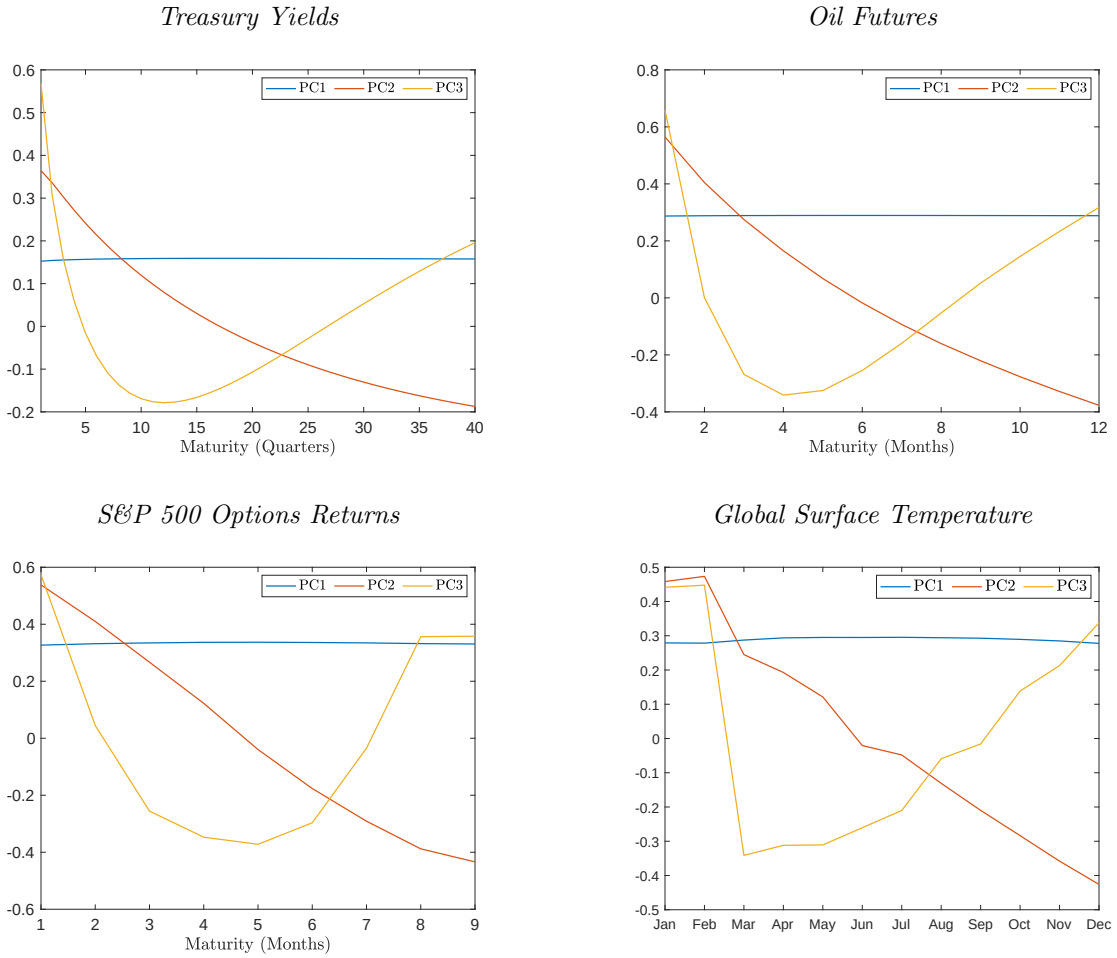
$$r_{t,t+1}^{(n)} \equiv p_{t+1}^{(n-1)} - p_t^{(n)}. \tag{3}$$

The corresponding excess return is then given by

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

[3] Appendix A provides full details about the data used in Figure 1.

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 A.



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 the notation to $r_{t+1}^{(n)}$ and $rx_{t+1}^{(n)}$, respectively. Observe that since returns are defined based on changes in prices, they directly inherit the overlapping nature of prices relative to forward rates. 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")

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

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

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

Similar to the arguments above, the overlapping sum that links difference and excess returns induces strong cross-sectional correlations which principal components analysis interprets as evidence of only a small number of underlying factors. Difference returns, then, have the appealing property that they have minimal time series persistence (unlike forwards and yields) and no mechanical cross-sectional persistence (unlike returns). As such, it is advantageous to base our bootstrap on difference returns as the main primitive object.

Further intuition for our bootstrap method can be gleaned from the following identity:

$$\begin{aligned} f_t^{(n)} &= f_t^{(n)} + f_{t-1}^{(n+1)} - f_{t-1}^{(n+1)} + \cdots + f_{t-N+n}^{(N)} - f_{t-N+n}^{(N)} && (7)\\ &= f_{t-N+n}^{(N)} - \left(f_{t-1}^{(n+1)} - f_t^{(n)}\right) - \left(f_{t-2}^{(n)} - f_{t-1}^{(n+1)}\right) - \cdots - \left(f_{t-N+n}^{(N)} - f_{t-N+n+1}^{(N-1)}\right) && (8)\\ &= f_{t-N+n}^{(N)} - dr_t^{(n+1)} - dr_{t-1}^{(n+2)} - \cdots - dr_{t-N+n+1}^{(N)}, && (9) \end{aligned}$$

where the transition from equation (8) to equation (9) utilizes the relationship between difference returns and forwards given by equation (5). Equation (9) clearly illustrates that difference returns, along with the longest horizon forward rate, serve as fundamental building blocks of the yield curve. More specifically, this equation implies that if we observe the longest horizon forward rate, $\left\{f_t^{(N)}\right\}_{t=2}^{T}$, all difference returns, $\left\{(dr_t^{(2)}, \ldots, dr_t^{(N)})\right\}_{t=2}^{T}$, along with an initial forward curve, $\left(f_1^{(1)}, \ldots, f_1^{(N)}\right)$, we are able to construct the entire forward curve $\left\{\left(f_t^{(1)}, \ldots, f_t^{(N)}\right)\right\}_{t=1}^{T}$ and yield curve $\left\{\left(y_t^{(1)}, \ldots, y_t^{(N)}\right)\right\}_{t=1}^{T}$.[4] Thus, conditional on an initial forward curve, we may jointly resample the N-maturity forward rate and the difference returns to generate bootstrap samples of the whole yield curve. A key feature of our bootstrap procedure is that we have chosen primitive objects in such a way that the only persistent object is the longest horizon forward rate which is confined to a single dimension and is also less persistent than other yields or forwards.

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

2.2 Bootstrap Method

We first describe our procedure for resampling the yield curve (hereafter referred to as the "CG bootstrap" for simplicity). 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 a matrix Z with its t-th row given by $z_t = (f_{t+1}^{(N)}, dr_{t+1}^{(2)}, ..., dr_{t+1}^{(N)})'$. Under some regularity conditions (stationarity and ergodicity) for the multivariate process z_t, the bootstrap samples $\{z_t^*\}$ for $t = 1, \ldots, T-1$ are then obtained by drawing with replacement blocks of $M < T$ rows (time-series dimension) from the matrix Z, jointly for all cross-sectional observations.[5] 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 $Z_{t,M} = (z_t, z_{t+1}, ..., z_{t+M-1})$ denote a block of M consecutive observations of z_t, $k = [(T-1)/M]$, where $[a]$ signifies the smallest integer that is greater than or equal to a, and define $\overline{T} = kM$.[6] For a given block size M, we then construct the blocks $\{Z_{t,M}\}_{t=1}^{\overline{T}/M}$. We resample with replacement k blocks from $(Z_{1,M},\ Z_{2,M},\ ...,\ Z_{T-M,M})$ by drawing k *i.i.d.* discrete uniform random variables $(u_1, ..., u_k)$ and obtain $Z_{u_1,M}$ through $Z_{u_k,M}$. Then, the bootstrap sample is given by $Z^* = [(z_1^*,\ z_2^*,\ ...,\ z_M^*),\ (z_{M+1}^*,\ z_{M+2}^*,\ ...,\ z_{2M}^*),\ ...,\ (z_{\overline{T}-M+1}^*,\ z_{\overline{T}-M+2}^*,\ ...,\ z_{\overline{T}}^*)]'$. Since the cross-sectional dimension N in term structure data is generally a nontrivial fraction of the sample size T, this informs our choice of M which is guided by a simple rule-of-thumb:

$$M = (TN)^{2/5}. \tag{10}$$

Given the bootstrap sample $\{f_t^{(N)*}, dr_t^{(2)*}, ..., dr_t^{(N)*}\}$ and the initial forward curve,[7] the bootstrap forward rates with maturities 1 to $N-1$ periods are constructed, using the identity given in

[5] 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.

[6] $\overline{T}$ is the smallest block bootstrap sample which has at least T observations. In practice, we select the first T observations of the block bootstrap sample.

[7] In all numerical results, we initialize the forward curve with the first sample observation. However, there may be circumstances where it is desirable to conduct inference conditional on a different initial condition. Our results accommodate such a case.

equation (5), as

$$f_{t+1}^{(N-1)*} = f_t^{(N)*} - dr_{t+1}^{(N)*} \tag{11}$$

$$f_{t+1}^{(N-2)*} = f_t^{(N-1)*} - dr_{t+1}^{(N-1)*} \tag{12}$$

$$\vdots$$

$$f_{t+1}^{(1)*} = f_t^{(2)*} - dr_{t+1}^{(2)*}, \tag{13}$$

for $t = 1, ..., T-1$. Finally, the bootstrap yields and returns are obtained, respectively, as $y_t^{(n)*} = \frac{1}{n}\sum_{i=1}^{n} f_t^{(i)*}$ and $rx_t^{(n)*} = \sum_{i=2}^{n} dr_t^{(i)*}$ for $n = 1, ..., N$ and $t = 2, ..., T$ by equations (2) and (6).

To visualize how well our resampling procedure mimics the actual dynamics of various yield curve-based variables, Figure 2 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.[8]

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. Let

$$rx_{t+h}^{(n,h)} \equiv p_{t+h}^{(n-h)} - p_t^{(n)} + p_t^{(h)} \tag{14}$$

be the h-period excess holding period return on a n-maturity bond. 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[9] and w_t being macro or other financial predictors. More explicitly, we focus on inference in the following predictive regression model

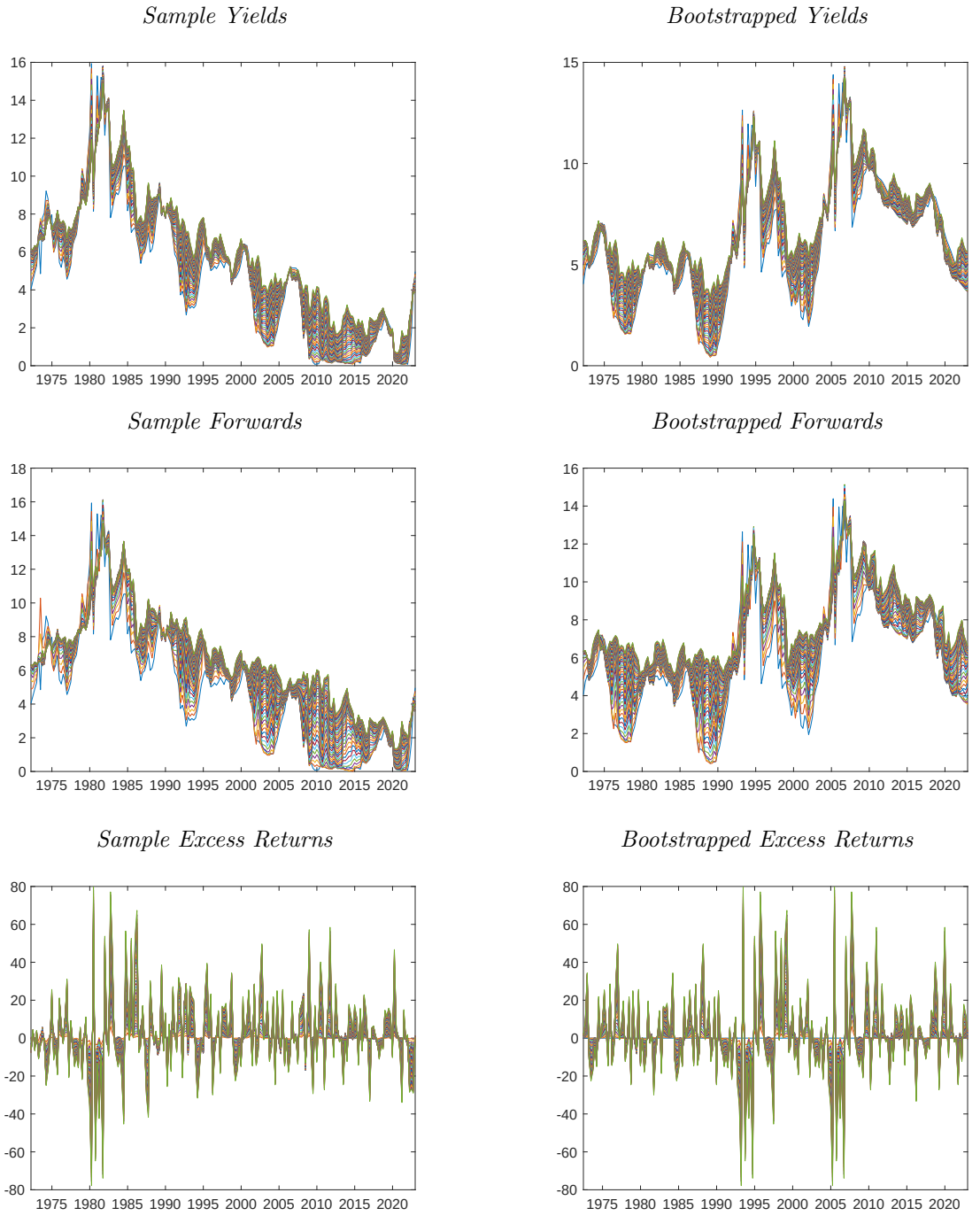
$$rx_{t+h}^{(n,h)} = \alpha^{(n)} + \beta^{(n)\prime} x_t + \varepsilon_{t+h}^{(n)} \tag{15}$$

$$= \alpha^{(n)} + \beta_1^{(n)\prime} g_t + \beta_2^{(n)\prime} w_t + \varepsilon_{t+h}^{(n)}, \tag{16}$$

[8] 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.

[9] For example, 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), or a linear combination of yields (e.g., Joslin et al. 2011) or forwards (e.g., Cochrane and Piazzesi 2005, 2008).

Figure 2. 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.



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.

To accommodate the case of external predictors, 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 $\tilde{z}_t = (f_{t+1}^{(N)}, dr_{t+1}^{(2)}, ..., dr_{t+1}^{(N)}, w_{t+1}')'$. 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 $y_t^{(n)*}$ or $f_t^{(n)*}$ which guarantees that all yield curve identities between variables hold in each bootstrapped sample. The bootstrapped external predictors are the last columns of $\tilde{Z}^*$ and are used to complete the full set of bootstrapped predictors $x_t^* = (g_t^{*\prime}, w_t^{*\prime})'$.

The bootstrap OLS estimator in the predictive regression is then given by

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

where $\bar{x}^* = (T-h)^{-1} \sum_{t=1}^{T-h} x_t^*$. Let $\hat{\Omega}_T$ denote the HAC estimator of the variance matrix of the OLS estimator $\hat{\beta}^{(n)}$ and $\hat{\Omega}_T^*$ be its bootstrap analog. With this notation, define the sample t-statistic for the ith predictor as $\hat{t}_i = (\hat{\beta}_i^{(n)} - \beta_i^{(n)})/\text{se}(\hat{\beta}_i^{(n)})$, where $\text{se}(\hat{\beta}^{(n)}) = \sqrt{\text{diag}(\hat{\Omega}_T)}$, and $t_i^* = (\hat{\beta}_i^{(n)*} - \hat{\beta}_i^{(n)})/\text{se}(\hat{\beta}_i^*)$, where $\text{se}(\hat{\beta}^{(n)*}) = \sqrt{\text{diag}(\hat{\Omega}_T^*)}$. The p-values for the symmetric percentile-t bootstrap method are obtained as

$$p_{|t_i|}^* = \frac{1}{B} \sum_{j=1}^{B} I\left\{ |t_{i,j}^*| > |\hat{t}_i| \right\}, \tag{17}$$

where $I\{\cdot\}$ denotes the indicator function, B is the number of bootstrap replications, and $t_{i,j}^*$ is t_i^* for the j-th ($j = 1, ..., B$) bootstrapped sample. These are the bootstrap p-values that we utilize in our simulations and empirical applications. The technical details regarding the validity of this bootstrap procedure are provided in the SA.

2.2.1 Implementational Details for the CG Bootstrap

Pre-whitening. While long-horizon forward rates tend to exhibit less persistence than the short rate, $f_t^{(N)}$ can still be very persistent depending on the frequency of the data and the sample 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 a similar degree of time-series dependence. In this case, we jointly "pre-whiten" 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. 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 $\hat{\Phi}$. These pre-whitened 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

$$\begin{pmatrix} f_t^{(N)*} \\ w_t^* \end{pmatrix} = \hat{\mu} + \hat{\Phi} \begin{pmatrix} f_{t-1}^{(N)*} \\ w_{t-1}^* \end{pmatrix} + \begin{pmatrix} \tilde{f}_t^{(N)*} \\ \tilde{w}_t^* \end{pmatrix} \tag{18}$$

for $t = 2, ..., T$, $f_1^{(N)*} = f_1^{(N)}$ and $w_1^* = w_1$.[10] We then proceed as discussed above.

Bias Correction. In all empirical applications and simulation experiments, we use the bootstrap bias correction of the coefficient matrix $\hat{\Phi}$ as suggested by 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). We only implement the bias correction when the maximum eigenvalue is less than one, as advocated in the bootstrap literature (e.g., Kilian 1998 and Kilian and Lütkepohl 2017, p.364). In our practical implementation, we use $1,000$ bootstrap replications when utilizing the Kilian (1998) bias correction. In our simulation experiments in the next section, we show that the difference between the bootstrap implemented with and without bias correction is small. However, our recommendation is to use bias-corrected estimates for the pre-whitening process as, in practice, the degree of persistence is unknown.

Overlapping Returns. 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 a larger-than-usual bandwidth/truncation

[10] As is standard in the bootstrap literature, we demean $\tilde{f}_t^{(N)*}$ and $\tilde{w}_t^*$ before re-whitening using equation (18).

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 our baseline choice of HAC estimator, we use the Newey and West (1987) estimator with lag length h. 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.

Gaps in Maturities. 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 which are available at monthly frequency in the time series but only for annual maturities in the cross-sectional dimension. In this case, we can follow a similar design to the procedure described above. Suppose we observe K prices, $\left\{\left(p_t^{(n_1)}, p_t^{(n_2)}, \ldots, p_t^{(n_K)}\right)\right\}_{t=1}^{T}$ where $N = n_K$. In the Fama-Bliss data, we would have $K = 5$ and $(n_1, n_2, n_3, n_4, n_5) = (12, 24, 36, 48, 60)$. The analog of $f_t^{(N)}$ in this setting is

$$z_{1t} = \left(p_t^{(n_{K-1})} - p_t^{(n_K)}\right).$$

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

$$\begin{aligned}
z_{2t} &= p_t^{(n_{K-2})} - p_t^{(n_{K-1})} - \left(p_{t-1}^{(n_{K-1})} - p_{t-1}^{(n_K)}\right) \\
z_{3t} &= p_t^{(n_{K-3})} - p_t^{(n_{K-2})} - \left(p_{t-1}^{(n_{K-2})} - p_{t-1}^{(n_{K-1})}\right) \\
&\vdots \\
z_{n_K t} &= p_t^{(0)} - p_t^{(n_1)} - \left(p_{t-1}^{(n_1)} - p_{t-1}^{(n_2)}\right) = -p_t^{(n_1)} - \left(p_{t-1}^{(n_1)} - p_{t-1}^{(n_2)}\right).
\end{aligned}$$

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

$$p_2^{(n_{K-2})*} - p_2^{(n_{K-1})*} = z_{22}^* + \left(p_1^{(n_{K-1})} - p_1^{(n_K)}\right),$$

and so on.

No Observed Short Rate. 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 Treasury TIPS market and the German Bund yield curve published by the Bundesbank (see Speck 2023 for further discussion) do not have a short-rate directly available. 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 (11)–(13) for this setting. In particular, we can simply replace the terminal recursion in equation (13) by

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

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

2.2.2 Discussion

There are three important aspects of our proposed bootstrap method. First, the building blocks of the CG bootstrap have weaker time-series persistence than yields or their principal components. However, our bootstrapped yields and forward rates will still exhibit high time-series persistence because they are generated via identities of the primitive objects that we resample.[11] As a result, the persistence of these predictors arises naturally from the relationships embedded in the term structure of interest rates.

Second, we bootstrap the entire cross-section of the yield curve which allows us to remain agnostic about the exact form of the strong cross-sectional dependence. Crump and Gospodinov (2022) argue that characterizing the true factor space in the term structure of interest rates is challenging and committing to a low-dimensional parametric structure can result in large hedging and portfolio allocation errors (see also Filipović et al. 2024). In contrast, our resampling method retains the underlying factor structure in the primitive processes and reconstructs the entire yield curve in an identity-preserving fashion.

Third, the CG bootstrap procedure is model free. Bond returns exhibit predictability using

[11] 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 2.1.

past yield and forward rate information. A standard parametric bootstrap procedure faces a number of difficulties in bootstrapping data in an internally consistent fashion that will replicate this predictability as well as other properties of the yield data. Our proposed bootstrap method ensures this internal consistency in mimicking the salient properties of the data by accommodating unknown forms of cross-sectional dependence and time-series persistence.

3 Simulation Evidence

In this section, we provide extensive simulation evidence demonstrating the performance of our new bootstrap procedure relative to both feasible and infeasible alternatives. In our simulation designs, forward rates are assumed to follow

$$F_t = \mathbf{a}^f + \mathbf{B}^f g_t + e_t, \tag{19}$$

where F_t is a vector of observed forward rates at time t, g_t denotes a subset of individual forward rates spanning the factor structure of the yield curve, and e_t is a vector of independent (over time) Gaussian measurement errors with a reduced-rank variance matrix, Σ_e.[12] As is standard in the literature, we assume that the factors, g_t, follow a VAR(1),

$$g_t = \mu_g + \Phi_g g_{t-1} + \eta_t^g, \tag{20}$$

where η_t^g are *i.i.d.* Gaussian errors with variance matrix Σ_η^g. To calibrate the parameters for our simulations, we use monthly data with selected forward rates as the factors g_t, and choose $F_t = (y_t^{(12)}, f_t^{(24,12)}, f_t^{(36,12)}, \ldots, f_t^{(108,12)}, f_t^{(120,12)})'$, where

$$f_t^{(n,h)} \equiv p_t^{(n-h)} - p_t^{(n)}. \tag{21}$$

For example, $f_t^{(24,12)}$ is the one-year rate, one year in the future and note that $f_t^{(n,1)} = f_t^{(n)}$ as defined in equation (1). We vary the choice of g_t to accommodate 3- and 5-factor specifications. All necessary parameters ($\mathbf{a}^f$, $\mathbf{B}^f$, Σ_e, μ_g, Φ_g, and Σ_η^g) are calibrated using system OLS estimates based on monthly Gurkaynak et al. (2007) data over the sample period 1972M1–2022M12 (612

[12]The variance matrix is of reduced rank to ensure internal consistency of the data-generating process, i.e., that the factors, g_t, are observed without error (e.g., Ang and Piazzesi 2003).

observations).

We first simulate F_t for $t = 1, \ldots, T$ using equations (19) and (20). We then obtain simulated prices, $p_t^{(12)}, p_t^{(24)}, ..., p_t^{(120)}$ using equation (2), and simulated excess holding period returns using equation (14). We run regressions of the form:

$$rx_{t+12}^{(n,12)} = \alpha^{(n)} + \beta_1^{(n)\prime} g_t + v_{g,t}, \tag{22}$$

and conduct inference on the elements of $\beta_1^{(n)}$. The values of the true parameters, $\beta_1^{(n)}$, are obtained analytically as

$$\beta_1^{(n)} = L\left(\Xi_1 \mathbf{B}^f - \Xi_2 \mathbf{B}^f \sum_{i=0}^{h-1} \Phi_g^h \mu_g\right), \tag{23}$$

where Ξ_1 and Ξ_2 denote the conformable identity matrix with the first row removed and last row removed, respectively, and L is a conformable lower triangular matrix of ones. For power computations, we consider tests of $\mathbb{H}_0 : \beta_{1i}^{(n)} = 0$, where $\beta_{1i}^{(n)}$ is the ith element of $\beta_1^{(n)}$. We omit results for the constant term as it is not generally a parameter of interest in these applications. In all simulations, we set $T = 600$, $B = 399$, and $S = 5000$. In the main text, we present results at 10% nominal level and in the SA, we provide the analogous results at 5% nominal level.

We report results for the following three resampling procedures:[13]

Oracle [O]. This option serves as an infeasible benchmark. The oracle inference procedure uses the full parametric structure of the data-generating process, including knowledge of the true parameters. The only sampling uncertainty in this method arises from the randomness generated from the B simulated draws of e_t and η_t^g. As $B \to \infty$, the oracle procedure, centered at the true parameter value, obtains the exact finite-sample distribution of the t-statistic. However, we fix B to be the same for all procedures to enable direct comparisons.

Bauer and Hamilton (2018) [BH]. We implement the BH bootstrap exactly as described in Bauer and Hamilton (2018). Let Y_t be the simulated yields obtained using F_t and equation (2).

[13] We also investigate the performance of 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.

Denote the first three principal components of yields as PC_t, where $\text{PC}_t = \hat{\Lambda}' Y_t$. Here, $\hat{\Lambda}$ is the 10×3 matrix of eigenvectors of the sample variance matrix of Y_t corresponding to the three largest eigenvalues and satisfying $\hat{\Lambda}'\hat{\Lambda} = I_3$. Define $\hat{\sigma}^2_{\text{PC}}$ as the estimator of the variance of the fitting error from principal components estimation,

$$\hat{\sigma}^2_{\text{PC}} = (10T)^{-1} \sum_{t=1}^{T} (Y_t - \hat{\Lambda}\hat{\Lambda}' Y_t)'(Y_t - \hat{\Lambda}\hat{\Lambda}' Y_t). \tag{24}$$

The dynamics of PC_t are estimated as a VAR(1) model resulting in

$$\text{PC}_t = \widehat{\mu}_{\text{PC}} + \widehat{\Phi}_{\text{PC}} \text{PC}_{t-1} + \widehat{\eta}^{\text{PC}}_t. \tag{25}$$

To obtain bootstrapped yield curves, we first set $\text{PC}^*_1 = \text{PC}_1$ and for $t > 1$ we use the recursions

$$\text{PC}^*_t = \widehat{\mu}_{\text{PC}} + \widehat{\Phi}_{\text{PC}} \text{PC}^*_{t-1} + \widehat{\eta}^{\text{PC}*}_t, \tag{26}$$

where $\widehat{\eta}^{\text{PC}*}_2, \ldots, \widehat{\eta}^{\text{PC}*}_T$ are independently sampled (with replacement) from the empirical distribution of $\widehat{\eta}^{\text{PC}}_t$. Then,

$$Y^*_t = \hat{\Lambda} \cdot \text{PC}^*_t + \hat{\sigma}_{\text{PC}} \cdot \nu_t, \tag{27}$$

where ν_t are *i.i.d.* Gaussian random vectors with variance matrix I_{10}. Bootstrapped forwards and returns can then be constructed using Y^*_t for $t = 1, \ldots, T$. In practice, yields (and consequently, principal components of yields) exhibit a high degree of persistence. It is often recommended that the parameter estimates in equation (25) be bias-adjusted (Bauer et al. 2012; Bauer et al. 2014). In the main text, we present results using bias-adjusted parameter estimates based on Kilian (1998) and report results without bias correction in the SA.[14]

[CG]. To implement our nonparametric bootstrap, we follow exactly the steps presented in Section 2.2. We use the block size choice given in equation (10) and employ the Kilian (1998) bias adjustment procedure for the pre-whitening process. We use the first observation in the simulated

[14] In order to exactly replicate the BH bootstrap, we follow the BH replication code which implements the bias correction for the OLS estimator of Φ_{PC} even when the maximal eigenvalue of $\widehat{\Phi}_{\text{PC}}$ is greater than or equal to one. As in BH, we use 1,000 bootstrap replications for the bias correction.

data as the initial condition for all bootstrapped samples.

Unlike the prior two procedures, our method does not aim to exploit any knowledge of the true data-generating process.

3.1 Simulations with Yield-Based Predictors

We first explore the finite-sample size and power properties of the three different resampling procedures in the presence of only yield-based predictors. The true parameters are $\beta_1^{(n)}$ as given in equation (23). Let $\hat{\beta}_{1i,s}^{(n)}$ be the OLS estimator for the sth simulation with corresponding t-statistic $\hat{t}_{1i,s}^{(n)}$ based on the Newey and West (1987) variance estimator with 12 lags. Similarly, let $\widehat{\beta}_{1i,s(b)}^{(n)*}$ and $\widehat{t}_{1i,s(b)}^{(n)*}$ be the bth bootstrap counterpart to $\widehat{\beta}_{1i,s}^{(n)}$ and $\widehat{t}_{1i,s}^{(n)}$, respectively, for the sth simulation. We calculate bootstrap p-values using the formula in equation (17). To obtain the bootstrap p-value under the null hypothesis, the test statistics $\widehat{t}_{1i,s}^{(n)}$ are centered at the true value of $\beta_{1i}^{(n)}$. To obtain the bootstrap p-value under the alternative hypothesis, the test statistics $\widehat{t}_{1i,s}^{(n)}$ are centered at zero. We then obtain empirical size (power) as the share of simulations where the bootstrap p-value constructed under the null (under the alternative) exceeds the pre-specified nominal level α.

We consider two specifications. For the first specification, we assume a three-factor model of the term structure and choose $g_t = (y_t^{(12)}, f_t^{(60,12)}, f_t^{(120,12)})'$. To provide context, in this design, the calibrated Φ_g matrix from the data has a maximum eigenvalue of .992 and all three eigenvalues are above 0.85. Note also that this setup aligns closely with the parametric structure which is imposed by the BH bootstrap. Table 1 presents the empirical size of the different methods for this design.

The BH bootstrap under-rejects relative to the nominal size of 10% 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 contrast, the empirical size for our nonparametric bootstrap is very close to the nominal size of 10% uniformly across maturities and for all three factors.[15] It is important to reiterate that the nonparametric bootstrap is completely agnostic about how the term structure data are generated. The performance by our method is even more impressive when compared to the infeasible oracle method whose rejection rates under the null hypothesis are very similar despite the fact that it uses full knowledge of the underlying DGP.

Table 2 reports the associated power in this simulation design. In terms of power, the rejection

[15] Results for a nominal size of 5% are qualitatively similar and presented in Section SA-2.3 in the SA.

Table 1. Size: Three-Factor VAR(1). This table presents empirical size for the three bootstrap methods: the nonparametric bootstrap introduced in Section 2.2 (CG); the parametric bootstrap of Bauer and Hamilton (2018) (BH); and the oracle bootstrap (O). The nominal level is 10% and the sample size is $T = 600$. Each column reports results for the t-test associated with the regressor $(g_{1t}, g_{2t}, g_{3t}) = (y_t^{(12)}, f_t^{(60,12)}, f_t^{(120,12)})$. Each row reports results for bond returns of the corresponding maturity. Based on 5,000 simulations and 399 bootstrap replications per simulation.

	CG Bootstrap			**BH Bootstrap**			**O Bootstrap**		
	g_{1t}	g_{2t}	g_{3t}	g_{1t}	g_{2t}	g_{3t}	g_{1t}	g_{2t}	g_{3t}
2y	0.107	0.100	0.079	0.045	0.023	0.013	0.097	0.101	0.100
3y	0.102	0.107	0.083	0.049	0.017	0.006	0.102	0.102	0.098
4y	0.102	0.114	0.086	0.051	0.011	0.001	0.104	0.101	0.098
5y	0.101	0.113	0.091	0.049	0.010	0.000	0.105	0.100	0.097
6y	0.103	0.113	0.090	0.049	0.006	0.000	0.104	0.099	0.098
7y	0.103	0.112	0.090	0.047	0.006	0.000	0.106	0.099	0.097
8y	0.104	0.112	0.090	0.046	0.006	0.000	0.107	0.100	0.096
9y	0.109	0.114	0.089	0.046	0.008	0.000	0.106	0.101	0.097
10y	0.110	0.112	0.088	0.045	0.005	0.000	0.108	0.105	0.096

Table 2. Power: Three-Factor VAR(1). This table presents empirical power for the three bootstrap methods: the nonparametric bootstrap introduced in Section 2.2 (CG); the parametric bootstrap of Bauer and Hamilton (2018) (BH); and the oracle bootstrap (O). The nominal level is 10% and the sample size is $T = 600$. Each column reports results for the t-test associated with the regressor $(g_{1t}, g_{2t}, g_{3t}) = (y_t^{(12)}, f_t^{(60,12)}, f_t^{(120,12)})$. Each row reports results for bond returns of the corresponding maturity. Based on 5,000 simulations and 399 bootstrap replications per simulation.

	CG Bootstrap			**BH Bootstrap**			**O Bootstrap**		
	g_{1t}	g_{2t}	g_{3t}	g_{1t}	g_{2t}	g_{3t}	g_{1t}	g_{2t}	g_{3t}
2y	0.262	0.236	0.079	0.254	0.160	0.005	0.307	0.236	0.109
3y	0.431	0.371	0.089	0.406	0.256	0.010	0.477	0.365	0.103
4y	0.552	0.436	0.100	0.513	0.303	0.006	0.589	0.428	0.109
5y	0.626	0.445	0.099	0.575	0.326	0.004	0.660	0.437	0.106
6y	0.673	0.422	0.092	0.614	0.288	0.000	0.696	0.411	0.098
7y	0.694	0.378	0.087	0.634	0.234	0.000	0.715	0.363	0.097
8y	0.697	0.326	0.093	0.636	0.171	0.000	0.724	0.305	0.111
9y	0.698	0.268	0.111	0.624	0.138	0.000	0.718	0.246	0.142
10y	0.688	0.213	0.152	0.615	0.060	0.000	0.707	0.202	0.191

rates for our bootstrap are highest for the first factor reaching a peak of about 70% for long maturities. The power for the second and third factor are noticeably lower. However, this is a feature of the simulation design, reflecting the relative proximity of the true parameter to zero, rather than a defect in the CG bootstrap. To see this, we can compare the power from the CG bootstrap to that of the oracle method (O bootstrap). The oracle bootstrap has power that is, at most, only modestly above that of the CG bootstrap. Since this benchmark method is both infeasible and uses all available information, the relatively small difference in power compared to our bootstrap confirms the excellent performance of our procedure. Note that for some entries, the power of the CG bootstrap exceeds that of the O bootstrap. This can arise from either slight

relative overrejections under the null hypothesis or some residual randomness from the limited simulation or bootstrap draws. Finally, we also report power for the BH bootstrap. This procedure has notably lower rejection rates than the CG bootstrap.

Table 3. Size: Five-Factor VAR(1). This table presents empirical size for the three bootstrap methods: the nonparametric bootstrap introduced in Section 2.2 (CG); the parametric bootstrap of Bauer and Hamilton (2018) (BH); and the oracle bootstrap (O). The nominal level is 10% and the sample size is $T = 600$. Each column reports results for the t-test associated with the regressor $(g_{1t}, g_{2t}, g_{3t}, g_{4t}, g_{5t}) = (y_t^{(12)}, f_t^{(36,12)}, f_t^{(60,12)}, f_t^{(84,12)}, f_t^{(120,12)})$. Each row reports results for bond returns of the corresponding maturity. Based on 5,000 simulations and 399 bootstrap replications per simulation.

	CG Bootstrap					**BH Bootstrap**					**O Bootstrap**				
	g_{1t}	g_{2t}	g_{3t}	g_{4t}	g_{5t}	g_{1t}	g_{2t}	g_{3t}	g_{4t}	g_{5t}	g_{1t}	g_{2t}	g_{3t}	g_{4t}	g_{5t}
2y	0.101	0.104	0.094	0.093	0.097	0.042	0.023	0.012	0.005	0.013	0.097	0.108	0.107	0.098	0.101
3y	0.100	0.098	0.097	0.091	0.093	0.030	0.013	0.013	0.008	0.013	0.098	0.109	0.108	0.103	0.099
4y	0.100	0.098	0.095	0.093	0.092	0.022	0.005	0.012	0.007	0.015	0.097	0.106	0.105	0.106	0.101
5y	0.101	0.095	0.095	0.094	0.092	0.019	0.005	0.011	0.010	0.015	0.097	0.104	0.104	0.107	0.100
6y	0.100	0.096	0.095	0.094	0.095	0.016	0.004	0.014	0.008	0.013	0.099	0.106	0.104	0.101	0.102
7y	0.102	0.099	0.095	0.095	0.096	0.015	0.003	0.014	0.007	0.012	0.099	0.108	0.106	0.103	0.101
8y	0.101	0.100	0.096	0.096	0.095	0.014	0.004	0.015	0.008	0.011	0.101	0.107	0.106	0.103	0.102
9y	0.101	0.098	0.097	0.096	0.095	0.015	0.008	0.015	0.008	0.014	0.101	0.106	0.106	0.105	0.103
10y	0.101	0.101	0.097	0.097	0.095	0.014	0.008	0.015	0.007	0.013	0.102	0.105	0.106	0.108	0.104

Table 4. Power: Five-Factor VAR(1). This table presents empirical power for the three bootstrap methods: the nonparametric bootstrap introduced in Section 2.2 (CG); the parametric bootstrap of Bauer and Hamilton (2018) (BH); and the oracle bootstrap (O). The nominal level is 10% and the sample size is $T = 600$. Each column reports results for the t-test associated with the regressor $(g_{1t}, g_{2t}, g_{3t}, g_{4t}, g_{5t}) = (y_t^{(12)}, f_t^{(36,12)}, f_t^{(60,12)}, f_t^{(84,12)}, f_t^{(120,12)})$. Each row reports results for bond returns of the corresponding maturity. Based on 5,000 simulations and 399 bootstrap replications per simulation.

	CG Bootstrap					**BH Bootstrap**					**O Bootstrap**				
	g_{1t}	g_{2t}	g_{3t}	g_{4t}	g_{5t}	g_{1t}	g_{2t}	g_{3t}	g_{4t}	g_{5t}	g_{1t}	g_{2t}	g_{3t}	g_{4t}	g_{5t}
2y	0.363	0.507	0.565	0.597	0.511	0.037	0.000	0.000	0.000	0.000	0.392	0.515	0.593	0.649	0.557
3y	0.404	0.438	0.484	0.546	0.507	0.095	0.000	0.000	0.000	0.000	0.426	0.446	0.516	0.602	0.557
4y	0.408	0.366	0.412	0.515	0.519	0.128	0.000	0.000	0.000	0.000	0.427	0.376	0.456	0.580	0.570
5y	0.407	0.323	0.384	0.519	0.538	0.127	0.000	0.000	0.000	0.000	0.423	0.327	0.431	0.582	0.596
6y	0.408	0.299	0.386	0.543	0.561	0.149	0.000	0.000	0.000	0.000	0.422	0.304	0.431	0.604	0.618
7y	0.415	0.290	0.404	0.570	0.580	0.146	0.000	0.000	0.000	0.000	0.426	0.298	0.446	0.627	0.637
8y	0.424	0.288	0.426	0.591	0.583	0.144	0.000	0.000	0.000	0.000	0.434	0.298	0.465	0.643	0.635
9y	0.430	0.296	0.443	0.598	0.564	0.136	0.000	0.000	0.000	0.000	0.440	0.298	0.477	0.654	0.612
10y	0.435	0.297	0.449	0.590	0.528	0.136	0.000	0.000	0.000	0.000	0.446	0.299	0.479	0.641	0.576

For the second simulation design, we increase the number of factors to five by adding $f_t^{(36,12)}$ and $f_t^{(84,12)}$ and re-calibrate the parameters using the actual data.[16] This change will induce misspecification for the BH bootstrap as its prescription is to utilize only three principal components of yields. However, this misspecification would not be easily detected as the percentage of variation (averaged across simulations) in yields explained by the first three principal components is 97.40%,

[16] It is important to note that since we use forward rates as factors when expanding the factor space relative to the three-factor model, we are incorporating additional persistent and cross-sectionally dependent variables. Since we calibrate all parameters to the data, this will naturally lead to changes in all of the true regression coefficients that, in part, determine the power envelope.

2.47%, and 0.12%, respectively.

Table 3 presents the empirical size for this specification. Again, our nonparametric bootstrap has excellent size properties while the BH bootstrap is severely undersized across all five factors. As would be expected, the oracle method has size that is very close to the nominal level across all factors and maturities. Table 4 reports the associated power in this simulation design. As in the three-factor design, the CG bootstrap has power that compares favorably to the infeasible oracle bootstrap across all coefficients and maturities.[17] The power of the BH bootstrap is severely compromised by the misspecification due to a strong estimation bias induced by the incorrect factor space. Although there is some very modest 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 Tables 3 and 4 and are presented in the SA (Tables SA.9 and SA.10 in the SA). The CG bootstrap continues to have empirical size close to the nominal level whereas the BH bootstrap is severely undersized. Moreover, the CG bootstrap has power that is similar to that of the infeasible O bootstrap while this form of misspecification further deteriorates the power properties of the BH bootstrap.

3.2 Simulations with External Predictors

We now provide further simulation evidence by considering return regressions featuring both yield-based variables and external predictors as in equation (16). We assume that the external predictors follow a VAR(1):

$$w_t = \mu_w + \Phi_w w_{t-1} + \eta_t^w, \qquad t = 1, \ldots, T, \tag{28}$$

where η_t^w are *i.i.d.* Gaussian errors with variance matrix Σ_η^w. The true coefficients associated with $\beta_1^{(n)}$ are as before whereas the coefficients associated with the variables w_t, $\beta_2^{(n)}$, are identically

[17] In the SA, we present results for a sample size of $T = 2,000$ which confirm that the empirical power approaches 100% with the sample size.

zero.

We choose two standard macroeconomic predictors: the 3-month percentage change in core CPI inflation and the 3-month percentage change in industrial production. We then calibrate the necessary parameters for the VAR(1) model using system OLS estimates over the same sample period as for yields. Note that these predictors are quite persistent with the maximum eigenvalue of the autoregressive matrix equal to 0.94. We report empirical size for the null hypothesis that each element in $\beta_2^{(n)}$ is equal to zero along with the empirical size for the three coefficients in $\beta_1^{(n)}$.

Our proposed bootstrap method is implemented as described in Section 2. We pre-whiten $f_t^{(N)}$ and w_t jointly using a VAR(1) specification. For the BH bootstrap, the external predictors are bootstrapped under the null hypothesis of no return predictability. It is important to note that the BH method proceeds under the null hypothesis, $\mathbb{H}_0 : \beta_n^{(2)} = 0$, which holds in this design. In contrast, our resampling procedure is agnostic about how the simulated data were generated and remains fully nonparametric.

Table 5. Size: Three-Factor VAR(1) with External Predictors. This table presents empirical size for the three bootstrap methods: the nonparametric bootstrap introduced in Section 2.2 (CG); the parametric bootstrap of Bauer and Hamilton (2018) (BH); and the oracle bootstrap (O). The nominal level is 10% and the sample size is $T = 600$. Each column reports results for the t-test associated with the regressor $(g_{1t}, g_{2t}, g_{3t}) = (y_t^{(12)}, f_t^{(60,12)}, f_t^{(120,12)})$ and the external predictors w_{1t} and w_{2t}. Each row reports results for bond returns of the corresponding maturity. Based on 5,000 simulations and 399 bootstrap replications per simulation.

	CG Bootstrap					**BH Bootstrap**					**O Bootstrap**				
	g_{1t}	g_{2t}	g_{3t}	w_{1t}	w_{2t}	g_{1t}	g_{2t}	g_{3t}	w_{1t}	w_{2t}	g_{1t}	g_{2t}	g_{3t}	w_{1t}	w_{2t}
2y	0.111	0.098	0.081	0.098	0.101	0.046	0.019	0.015	0.107	0.106	0.096	0.103	0.100	0.101	0.107
3y	0.109	0.105	0.085	0.099	0.104	0.046	0.016	0.006	0.103	0.106	0.100	0.102	0.097	0.101	0.109
4y	0.104	0.110	0.086	0.103	0.105	0.047	0.009	0.002	0.103	0.107	0.100	0.101	0.098	0.102	0.109
5y	0.100	0.110	0.087	0.105	0.106	0.046	0.011	0.001	0.103	0.104	0.101	0.100	0.098	0.100	0.110
6y	0.099	0.110	0.085	0.107	0.106	0.046	0.007	0.000	0.102	0.104	0.103	0.097	0.096	0.102	0.107
7y	0.100	0.109	0.088	0.105	0.107	0.043	0.006	0.000	0.103	0.103	0.101	0.096	0.097	0.099	0.102
8y	0.100	0.110	0.086	0.107	0.104	0.043	0.006	0.000	0.104	0.103	0.100	0.096	0.098	0.101	0.102
9y	0.101	0.110	0.085	0.106	0.102	0.042	0.009	0.000	0.107	0.106	0.103	0.097	0.098	0.101	0.101
10y	0.100	0.110	0.085	0.108	0.103	0.042	0.006	0.000	0.105	0.105	0.104	0.099	0.098	0.101	0.103

Table 5 presents results for inference on the coefficients $\beta_1^{(n)}$ and $\beta_2^{(n)}$. For the yield-based predictors, the BH bootstrap continues to be severely undersized as in Table 1 whereas the CG bootstrap and the O bootstrap control size very well across maturities and the three different predictors. For the external predictors, both feasible methods (BH and CG bootstraps) work very well with empirical size very close to the 10% nominal level across maturities. We emphasize that the overlapping nature of these returns (since $h = 12$) produces strong serial correlation in the left-hand side variable.[18] With predictors which are also persistent, minimizing size distortions becomes

[18] For a heuristic example, note that an h-period moving average of a white noise process has first-order autocor-

Table 6. Power: Three-Factor VAR(1) with External Predictors. This table presents empirical power for the three bootstrap methods: the nonparametric bootstrap introduced in Section 2.2 (CG); the parametric bootstrap of Bauer and Hamilton (2018) (BH); and the oracle bootstrap (O). The nominal level is 10% and the sample size is $T = 600$. Each column reports results for the t-test associated with the regressor $(g_{1t}, g_{2t}, g_{3t}) = (y_t^{(12)}, f_t^{(60,12)}, f_t^{(120,12)})$. Each row reports results for bond returns of the corresponding maturity. Based on 5,000 simulations and 399 bootstrap replications per simulation.

	CG Bootstrap			**BH Bootstrap**			**O Bootstrap**		
	g_{1t}	g_{2t}	g_{3t}	g_{1t}	g_{2t}	g_{3t}	g_{1t}	g_{2t}	g_{3t}
2y	0.230	0.242	0.075	0.232	0.166	0.004	0.272	0.238	0.107
3y	0.403	0.375	0.089	0.382	0.262	0.011	0.446	0.367	0.102
4y	0.527	0.441	0.103	0.495	0.318	0.009	0.570	0.437	0.112
5y	0.611	0.454	0.100	0.562	0.338	0.005	0.648	0.447	0.111
6y	0.651	0.433	0.094	0.601	0.301	0.001	0.687	0.418	0.103
7y	0.676	0.386	0.082	0.619	0.247	0.000	0.706	0.371	0.097
8y	0.687	0.330	0.089	0.621	0.183	0.000	0.710	0.309	0.111
9y	0.684	0.274	0.108	0.618	0.142	0.000	0.706	0.250	0.142
10y	0.678	0.219	0.143	0.612	0.067	0.000	0.698	0.201	0.192

increasingly difficult. Despite this challenging predictive regression design, both feasible 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-2.4 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, Table 6 only reports power for the yield-based predictors. Despite the addition of two extraneous regressors, the power of the CG bootstrap is similar to that of Table 2 and comparable to the power of the O bootstrap.

Taken in sum, these simulation experiments show that it is preferable to avoid committing to a tightly parameterized, finite-dimensional factor structure as this may result in the omission of important information embedded in the yield curve.

4 Empirical Applications

4.1 Bond Returns and Economic Trends

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 and find that the inclusion of this variable substantially

relation coefficient of $\frac{h-1}{h}$.

improves upon bond return predictability relative to using yield-based information. Building off of Cieslak and Povala (2015), Bauer and Rudebusch (2020) ("BR," hereafter) use the BH bootstrap to argue that, in addition to trend inflation, bond returns are also predictable by the equilibrium real interest rate. In this section, we utilize our bootstrap to re-examine the evidence of bond return predictability based on these slow-moving economic fundamentals. Importantly, we show that our bootstrap can be tailored perfectly to this setting whereas the BH bootstrap used by BR is invalid. The source of this invalidity is that the BH bootstrap cannot accommodate predictors, outside of the the first three principal components of yields, which are themselves a function of the yield curve.

In the first part of this empirical application, we use the same data as in BR (see Appendix A for full description of the data and additional implementation details). The predictors are the first three (scaled) principal components, a measure of trend inflation, π_t^*, and a measure of the equilibrium real short-term interest rate, r_t^*. As a proxy for trend inflation, BR use the perceived target rate (PTR) which is an input in the FRB/US model (see BR for further details). BR present results for different proxies of the equilibrium real rate. We use their "moving average" measure, which is an exponential weighted moving average (EWMA) of the difference between the 3-month interest rate and year-over-year core PCE inflation. We focus on this proxy as it has the smallest reported p-values in BR.

We consider the model,

$$rx_{t+h}^{(n)} = \alpha + \beta_1^{(n)\prime} g_t + \beta_2^{(n)\prime} w_t + \varepsilon_t^{(n)}, \tag{29}$$

where $g_t = \text{PC}_t$ and w_t is one of three different specifications: (I) $w_t = \pi_t^*$, (II) $w_t = (\pi_t^*, r_t^*)'$, or (III) $w_t = r_t^*$. BR consider specifications (I) and (II) and we add specification (III) to investigate the marginal predictive power of r_t^*. We follow BR and construct (approximate) one-quarter returns when $h = 1$ and (exact) four-quarter returns when $h = 4$ for annual maturities of two years to 15 years.[19]

In Table 7, we exactly replicate the results for specifications (I) and (II) shown in Table 2 on p.1329 of BR for both samples they considered (1971Q4–2018Q1 and 1985Q1–2018Q1). The

[19]We follow BR and calculate one-quarter returns using the approximation described in Bauer and Rudebusch (2020, Footnote 21).

left-hand side variable is the cross-sectional average (across maturities) of one-quarter ahead bond returns. As in BR, standard errors are constructed using the Eicker-White variance estimator. Using this variance estimator, the magnitude of the t-statistics associated with π_t^* are above 4 and the t-statistic associated with r_t^* is 2.6. BR also report results for the restricted sample starting in 1985 which we show in the middle panel of Table 7. In this subsample, the magnitude of the t-statistic associated with r_t^* increases further, rising to 3.8. Across both of these samples, the addition of r_t^* as a predictor increases the regression R^2 notably, as shown in the bottom two rows of the table. For completeness, we also report the four-quarter ahead results for both of these samples using Newey-West standard errors. At the four-quarter ahead horizon, the magnitudes of the t-statistics continue to be very large.

Table 7. Replication of Bauer and Rudebusch (2020). This table replicates Table 2 in Bauer and Rudebusch (2020) along with an additional subsample (1971Q4-2007Q4) for the regression model given by equation (29). The dependent variable is the cross-sectional average of bond returns for annual maturities from two to 15 years and the independent variables are the first three principal components of yields, the perceived target rule (π^*), and an exponentially weighted moving average of real interest rates (r^*). We report OLS estimates with corresponding standard errors in parentheses. Standard errors are constructed with the Eicker-White variance matrix for one-quarter ahead returns and the Newey-West variance estimator with six lags for four-quarter ahead returns. The last two rows present the unadjusted and adjusted regression R^2 statistics.

	1971Q4–2018Q1				**1985Q1–2018Q1**				**1971Q4–2007Q4**			
	1-Qtr. Ahead		*4-Qtrs. Ahead*		*1-Qtr. Ahead*		*4-Qtrs. Ahead*		*1-Qtr. Ahead*		*4-Qtrs. Ahead*	
	(I)	(II)	(I)	(II)	(I)	(II)	(I)	(II)	(I)	(II)	(I)	(II)
PC1	0.98	2.04	2.70	5.34	0.59	2.38	1.80	6.77	1.72	2.05	4.96	5.38
	(0.26)	(0.56)	(0.69)	(0.99)	(0.22)	(0.51)	(0.52)	(1.01)	(0.43)	(0.62)	(0.80)	(0.96)
PC2	0.47	0.68	1.76	2.26	0.50	0.65	1.71	2.11	0.48	0.58	1.87	1.99
	(0.17)	(0.15)	(0.29)	(0.28)	(0.16)	(0.15)	(0.34)	(0.28)	(0.18)	(0.17)	(0.29)	(0.29)
PC3	-1.79	-0.90	-1.96	0.03	-0.97	2.11	-1.63	6.53	-2.88	-2.45	-4.20	-3.65
	(1.27)	(1.43)	(1.71)	(2.05)	(1.12)	(1.55)	(2.38)	(2.93)	(1.46)	(1.63)	(2.05)	(2.33)
π^*	-1.95	-3.89	-5.68	-10.52	-1.05	-3.34	-4.03	-10.36	-2.71	-3.48	-7.87	-8.86
	(0.44)	(0.92)	(1.17)	(1.80)	(0.73)	(0.83)	(1.36)	(1.31)	(0.54)	(1.02)	(1.31)	(1.78)
r^*		-2.71		-6.90		-4.11		-11.78		-1.45		-1.84
		(1.04)		(1.82)		(1.08)		(2.47)		(1.33)		(1.94)
R^2	0.16	0.20	0.43	0.51	0.10	0.18	0.31	0.53	0.23	0.24	0.58	0.59
adj. R^2	0.14	0.18	0.41	0.49	0.07	0.15	0.29	0.51	0.21	0.21	0.57	0.57

In the right panel of Table 7, we provide results for the pre-crisis sample 1971Q4–2007Q4. Compared to the other samples, the magnitude of the t-statistic associated with r_t^* declines sharply to about one while the t-statistic associated with π_t^* remains large. Furthermore, in the pre-crisis sample, adding r_t^* to the regression has no effect on the adjusted R^2. These latter results suggest that the strong predictive ability of r_t^* for future bond returns reported in BR is likely driven by the post-crisis sample. However, there are two important considerations when interpreting the results presented in Table 7. First, as discussed in Section 3, and shown in Section SA-2.4 of the SA,

HAC estimators, such as Newey-West, can lead to over-rejections of the null of no-predictability in settings with persistent predictors and (possibly) overlapping returns. Second, we may have cherry-picked the pre-crisis sample which fails to find a predictive role for r_t^* at either forecast horizon. In general, valid resampling-based procedures can circumvent both of these problems. In particular, an appealing aspect of bootstrap-based inference is that if the in-sample results are driven by a small number of observations (as appears to be the case here), then these observations will occur relatively infrequently in the bootstrapped samples. This leads to larger bootstrap-based critical values and wider bootstrap-based confidence intervals, making rejection of the null of no-predictability less likely.

Table 8. Bootstrap Inference on Return Predictability. This table provides bootstrap p-values for the coefficients of the regression model given by equation (29) based on the Bauer and Hamilton (2018) (BH) procedure and the nonparametric bootstrap introduced in Section 2.2 (CG). For reference, the OLS estimates for each specification are also shown. All results are based on 5,000 bootstrap replications.

	1971Q4–2018Q1						**1985Q1–2018Q1**					
	1-Qtr. Ahead			*4-Qtrs. Ahead*			*1-Qtr. Ahead*			*4-Qtrs. Ahead*		
	(I)	(II)	(III)	(I)	(II)	(III)	(I)	(II)	(III)	(I)	(II)	(III)
Est. Coeffs.												
π^*	-1.95	-3.89		-5.68	-10.52		-1.05	-3.34		-4.03	-10.36	
r^*		-2.71	0.68		-6.90	2.46		-4.11	-1.14		-11.78	-2.43
BH (p-val)												
π^*	0.00	0.00		0.01	0.00		0.36	0.00		0.14	0.00	
r^*		0.03	0.28		0.02	0.24		0.01	0.47		0.03	0.65
CG (p-val)												
π^*	0.02	0.25		0.02	0.13		0.50	0.10		0.15	0.03	
r^*		0.83	0.40		0.74	0.36		0.72	0.93		0.79	0.94

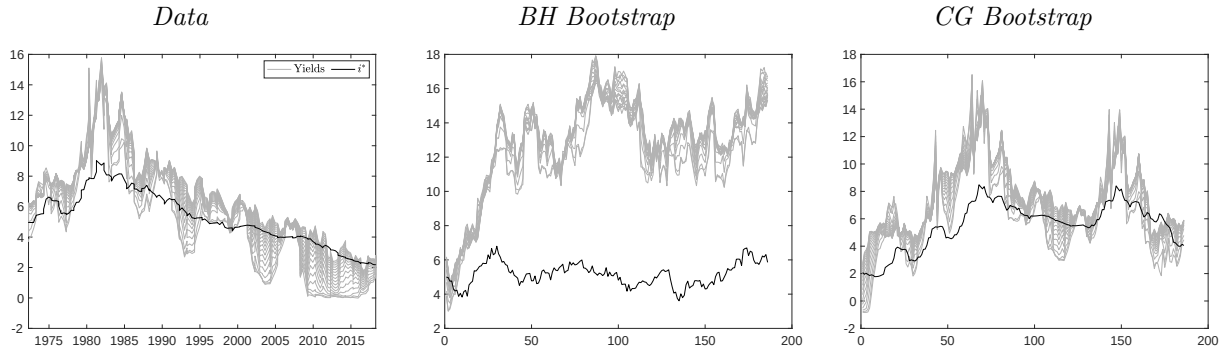
Using the BH bootstrap procedure, BR report strong, statistically significant predictive power of r_t^*. These results are reproduced in Table 8.[20] As in BR, the BH bootstrap produces p-values for the coefficients associated with both π_t^* and r_t^* that are generally close to zero for specification (II). However, when r_t^* is included as a predictor on its own in specification (III), the sign of the coefficient switches in the full sample and the BH bootstrap p-values are far away from 10% for both sample periods and forecast horizons. This is also the case for the CG bootstrap, shown in the bottom rows of Table 8, where, again, there is no evidence of marginal predictive power for r_t^*. In contrast to the BH bootstrap, the CG bootstrap also fails to find evidence for the predictive power of r_t^* in specification (II). Across all samples and forecast horizons the p-value is above 0.72. Taken in sum, our bootstrap results shown in Table 8 provide strong evidence against r_t^* as a predictor

[20]Note that there will be some differences in the p-values in Table 8 as compared to BR because of the inherent randomness of the bootstrap.

of bond returns.

What drives the stark differences in inference between the BH and CG bootstrap? Most importantly, the BH bootstrap *cannot* be used to conduct inference on measures of r_t^* as an external predictor. This is because the BH bootstrap is conducted under the null hypothesis that returns are only predictable based on lagged yields and any additional variables do not predict returns. This necessarily excludes any additional variables which, themselves, are a function of yields such as the proxy for r_t^* shown in Table 7. However, all of the r_t^* proxies considered in BR include the short-term interest rate as an input. This dependence is not reflected in the BH bootstrap as it fails to correctly incorporate the underlying structure and is consequently invalid. To see this clearly, Figure 3 shows realized yields and $i_t^* = \pi_t^* + r_t^*$ along with the same objects for a single bootstrap draw from the BH bootstrap and the CG bootstrap. Clearly, the i_t^* from the BH bootstrap does not behave as if it is the underlying trend in the corresponding bootstrapped yields. Yields have a strong upward trend that is entirely divorced from the bootstrapped i_t^* which comes about because the bootstrapped yields and i_t^* are decoupled in the BH procedure. In contrast, our bootstrap has the desirable feature that yields and i_t^* exhibit strong co-movement as shown in the right chart of Figure 3.[21]

Figure 3. Actual and Bootstrapped Yields and i_t^*. This figure shows actual and bootstrapped yields along with i_t^*. The left chart shows realized yields along with i_t^* constructed as the sum of the perceived target rule (π_t^*) and the exponentially weighted moving average of real interest rates (r_t^*). The middle and right chart show the bootstrap counterparts of yields and i_t^* based on the Bauer and Hamilton (2018) (BH) bootstrap and the nonparametric bootstrap introduced in Section 2.2 (CG).



The results in Tables 7 and 8 rely on the PTR as the measure of trend inflation. However, the PTR is an amalgamation of both model-based and survey-based measures of the trend in inflation. In contrast, Cieslak and Povala (2015) (hereafter, "CPo") introduce a simple and transparent

[21] Note that neither our bootstrap procedure nor the BH bootstrap ensure that resampled nominal yields remain positive.

measure of trend inflation that can be nested easily into our bootstrap procedure. Furthermore, the CPo measure of trend inflation is available in real-time which is not the case for the PTR. The measure of π_t^* introduced by CPo is defined as a truncated exponentially weighted moving average (EWMA) of year-over-year core inflation,

$$\pi^*_{\text{CPo},t} \equiv \frac{1-\phi}{1-\phi^{(H-1)}} \sum_{s=0}^{H-1} \phi^s \pi^{\text{yoy}}_{t-s}, \tag{30}$$

where π^{yoy}_{t-s} is year-over-year core CPI inflation lagged one month to reflect the timing of the data release. CPo use monthly data and choose $(\phi, H) = (0.987, 120)$. To adapt (ϕ, H) to the quarterly setting of BR, we use the modified parameters $(\phi, H) = (0.987^3, 40)$. Since the core CPI index is only revised in terms of seasonal factors, the year-over-year growth rate is (essentially) unrevised.[22] To ensure that our measure of r_t^* is also real time, we construct it using year-over-year core CPI inflation (lagged one month) instead of core PCE inflation as the latter can be revised substantially (Audoly et al. 2023). This also has the advantage of reducing the dimension of the VAR process for pre-whitening from three to two. To bootstrap the data in this setting, we follow the approach given in Section 2 and utilize a bivariate VAR(1):

$$\begin{pmatrix} f_t^{(N)} \\ \pi_t^{\text{yoy}} \end{pmatrix} = \hat{\mu} + \hat{\Phi} \begin{pmatrix} f_{t-1}^{(N)} \\ \pi_{t-1}^{\text{yoy}} \end{pmatrix} + \begin{pmatrix} \tilde{f}_t^{(N)} \\ \tilde{\pi}_t^{\text{yoy}} \end{pmatrix}. \tag{31}$$

Table 9 presents the results for this exercise using these real-time measures of π_t^* and r_t^*. The column titles denote the specification and the coefficient of interest: (I,π^*) and (II,π^*) are the coefficients associated with π_t^* in specifications (I) and (II), respectively, and (II,r^*) is the coefficient associated with r_t^* in specification (II). We consider the same sample periods and forecast horizons as in Table 8. We report p-values for regressions using bond returns for individual maturities with the last row representing average bond returns. Table 9 convincingly shows that there is no evidence that r_t^* has predictive content for future bond returns. Across all maturities, sample periods, and forecast horizons, the p-values for the coefficient associated with r_t^* are no lower than 0.59. If we restrict our attention to the one-quarter ahead forecast horizon, which is what is reported in BR, the p-values range from 0.68 to 0.87. This reinforces the conclusions drawn from Table 8 and

[22] To our knowledge, the only exception occurred in September 2000 due to an error in calculation of the price index (see The Wall Street Journal 2000).

Table 9. Bootstrap Inference with Real-Time Trend Estimates. This table provides bootstrap p-values for the coefficients of the regression model given by equation (29) based on the nonparametric bootstrap introduced in Section 2.2 (CG). The dependent variable is bond returns for a specific annual maturity (or average across maturities) and the independent variables are the first three principal components of yields, the Cieslak and Povala (2015) factor, and an exponentially weighted moving average of real interest rates based on core CPI inflation (r^*). All results are based on 5,000 bootstrap replications.

	1971Q4–2018Q1						**1985Q1–2018Q1**					
	1-Qtr. Ahead			*4-Qtrs. Ahead*			*1-Qtr. Ahead*			*4-Qtrs. Ahead*		
	(I,π^*)	(II,π^*)	(II,r^*)	(I,π^*)	(II,π^*)	(II,r^*)	(I,π^*)	(II,π^*)	(II,r^*)	(I,π^*)	(II,π^*)	(II,r^*)
2-yr	0.32	0.38	0.87	0.23	0.22	0.95	0.16	0.21	0.82	0.15	0.14	0.90
3-yr	0.26	0.32	0.84	0.14	0.15	0.86	0.15	0.17	0.84	0.12	0.12	0.90
4-yr	0.20	0.27	0.82	0.09	0.11	0.81	0.14	0.15	0.84	0.11	0.10	0.89
5-yr	0.17	0.22	0.80	0.06	0.09	0.77	0.13	0.12	0.84	0.09	0.08	0.87
6-yr	0.14	0.18	0.78	0.05	0.07	0.74	0.12	0.11	0.82	0.07	0.06	0.84
7-yr	0.12	0.16	0.76	0.04	0.06	0.71	0.11	0.09	0.80	0.06	0.05	0.81
8-yr	0.11	0.14	0.75	0.04	0.05	0.69	0.10	0.07	0.79	0.05	0.04	0.78
9-yr	0.10	0.12	0.74	0.03	0.04	0.67	0.09	0.06	0.77	0.04	0.03	0.74
10-yr	0.09	0.11	0.72	0.04	0.04	0.65	0.08	0.05	0.76	0.04	0.02	0.70
11-yr	0.09	0.10	0.71	0.04	0.04	0.63	0.08	0.05	0.74	0.04	0.02	0.67
12-yr	0.08	0.09	0.70	0.04	0.04	0.62	0.08	0.04	0.73	0.04	0.02	0.64
13-yr	0.07	0.09	0.69	0.04	0.04	0.61	0.08	0.04	0.72	0.04	0.01	0.62
14-yr	0.07	0.09	0.69	0.04	0.03	0.60	0.08	0.03	0.71	0.04	0.01	0.61
15-yr	0.06	0.08	0.68	0.04	0.03	0.59	0.08	0.03	0.71	0.04	0.01	0.60
Avg.	0.09	0.12	0.73	0.04	0.05	0.66	0.09	0.05	0.76	0.04	0.02	0.73

confirms that they hold across the maturity spectrum.

Table 9 shows that trend inflation, on the other hand, has predictive content beyond information in the current yield curve for future bond returns. We omit the estimated coefficients in this table but note that the estimated coefficient associated with π_t^* is negative in sign and increasing in magnitude with maturity across all specifications, forecast horizons, and sample periods. For the full sample period and one-quarter ahead prediction, we find that the bootstrap p-values associated with trend inflation are at 10% or below for longer maturities. The results are stronger for the post-1985 subsample and for the four-quarter ahead horizon. In fact, at the four-quarter ahead horizon, and across both sub-samples, p-values uniformly fall below 5% beyond the eight-year maturity. Furthermore, average bond returns (last row) have p-values of 5% or lower for both specifications and both sample periods for this longer forecast horizon. One possible explanation for why the results for trend inflation are stronger in the shorter sample is the more complicated dynamics arising from multiple hump-shaped periods in the inflation process in the 1970s whereas the shorter sample is dominated by the downward trend in inflation and yields.

While, at first glance, it may appear that the p-values for trend inflation in Table 9 are not as small as frequently encountered in the literature, it is important to emphasize that our bootstrap

procedure is accommodating all of the possible sources of uncertainty. It takes into consideration the uncertainty arising from the exponential smoothing, the high persistence of inflation, and the joint dynamics of interest rates and inflation. Furthermore, it should not be surprising that the p-values for the coefficient on trend inflation sometimes increase with the inclusion of r_t^* as a predictor. This would be the case even when including a completely independent, highly persistent process as it can contaminate estimation and inference on the coefficient associated with π_t^* (see Section 1 of Bauer and Hamilton (2018) for a discussion of this phenomenon).

To summarize, the results from Tables 7–9 are supportive of the role of trend inflation as a bond risk factor, as argued by Cieslak and Povala (2015), as future returns appear to be driven by deviations of the yield curve from a time-varying reference point determined by inflation (Rebonato and Hatano 2022). Finally, note that we have focused on the EWMA measure of r_t^* in BR but our bootstrap can handle any of the individual r_t^* proxies (or the average thereof) used in BR. More generally, our bootstrap can preserve the underlying structure and appropriately accommodate uncertainty from generated regressors that might arise, for example, from reduced-form regressions (e.g., Cooper and Priestley 2008), state-space models, machine learning methods (e.g., Bianchi et al. 2020) or structural models.

4.2 Regression-Based Tests of the Expectations Hypothesis

The flexibility of our bootstrap facilitates its use, without modification, in more general contexts. As an illustrative example, we investigate the sampling properties of regression-based tests of the expectations hypothesis (EH). It has been well established in the literature that the results of these tests can vary dramatically 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,[23] 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

[23] We follow the time-series econometrics literature and define an unbalanced regression as a specification where the dependent and explanatory variables are characterized by meaningfully different degrees of persistence and variability.

versions of the regression-based tests of the EH. More specifically, we nonparametrically resample the entire yield curve, rather than the specific maturities used in a particular specification, which allows us to make comparisons across all of these formulations in a unified way. By contrast, any parametric resampling approach would necessarily be model-specific imposing the expectations hypothesis or its failure.

We begin by showing that 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). To our knowledge, this result is new to the literature. By the definition of forward rates and multi-period holding returns, we have that

$$f_{t+m}^{(n-m)} = p_{t+m}^{(n-m-1)} - p_{t+m}^{(n-m)} + f_t^{(n)} - f_t^{(n)} \tag{32}$$

$$= p_{t+m}^{(n-m-1)} - p_{t+m}^{(n-m)} + f_t^{(n)} - \left(p_t^{(n-1)} - p_t^{(n)}\right) \tag{33}$$

$$= f_t^{(n)} - \left(rx_{t+m}^{(n,m)} - rx_{t+m}^{(n-1,m)}\right), \tag{34}$$

where the first equality follows by equation (1) and the last equality follows by equation (14). Thus, under rational expectations, a regression-based implementation of a test of the EH can be based on the specification:

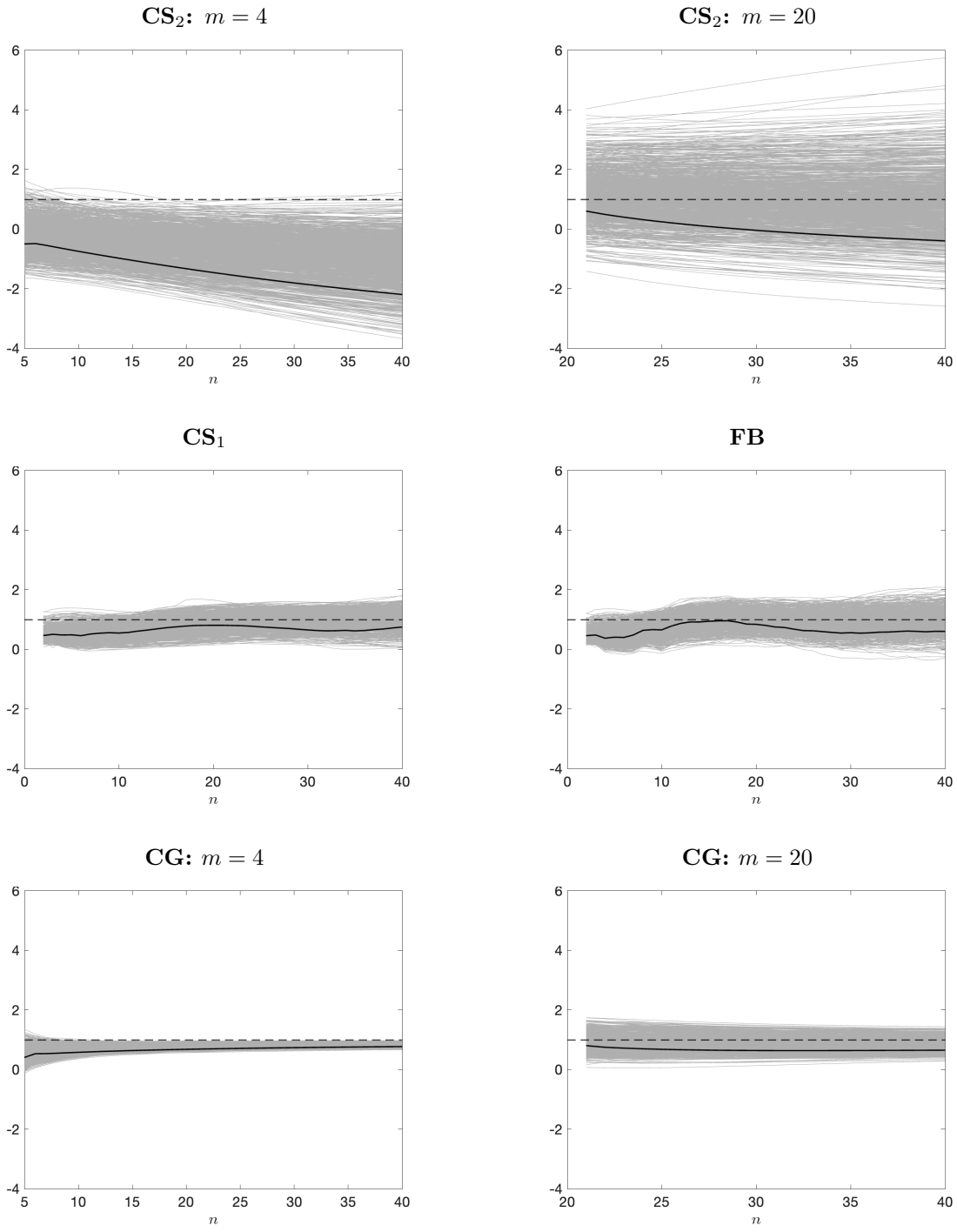
$$f_{t+m}^{(n-m)} = \alpha + \beta \cdot f_t^{(n)} + \epsilon_{t+m}^{(n,m)}. \tag{35}$$

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

$$f_{t+m}^{(n-m)} - y_t^{(1)} = \alpha + \beta \cdot \left(f_t^{(n)} - y_t^{(1)}\right) + \epsilon_{t+m}^{(n,m)}. \tag{36}$$

We refer to this regression-based test of the EH as the CG test. Equation (36) forms the basis for all existing tests of the EH; specifically, under the null hypothesis, any existing test can be obtained through linear combinations of equation (36) 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 (36) in the

Figure 4. Expectations Hypothesis Tests: OLS Sampling Properties. This figure presents the in-sample OLS estimate based on the realized data (black line) and 1,000 draws (grey lines) from the nonparametric bootstrap introduced in Section 2.2 (CG). The top row presents the estimates 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 black dashed horizontal line is at a value of one which corresponds to the EH. The sample period is 1972:Q1–2024:Q1.



special case where $m = n - 1$, i.e.,

$$y_{t+n-1}^{(1)} - y_t^{(1)} = \alpha + \beta \cdot \left(f_t^{(n)} - y_t^{(1)} \right) + \epsilon_{t+n-1}^{(n,n-1)}. \tag{37}$$

Furthermore, a time-series average of equation (37) gives

$$\frac{1}{n}\sum_{s=1}^{n} y_{t+s-1}^{(1)} - y_t^{(1)} = \alpha + \beta \cdot \left(\frac{1}{n}\sum_{s=1}^{n} f_t^{(s)} - y_t^{(1)}\right) + \frac{1}{n}\sum_{s=1}^{n} \epsilon_{t+s-1}^{(s,s-1)}, \tag{38}$$

and using equation (2) we obtain the first test of Campbell and Shiller (1991), denoted by CS_1,

$$\frac{1}{n}\sum_{s=0}^{n-1}\left(y_{t+s}^{(1)} - y_t^{(1)}\right) = \alpha + \beta\left[y_t^{(n)} - y_t^{(1)}\right] + v_{\mathrm{CS}_1,t+n-1}^{(n)}, \tag{39}$$

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 by CS_2, is

$$y_{t+m}^{(n-m)} - y_t^{(n)} = \alpha + \beta\frac{m}{n-m}\left(y_t^{(n)} - y_t^{(m)}\right) + v_{\mathrm{CS}_2,t+n-1}^{(n)}. \tag{40}$$

This can be obtained by taking a cross-sectional average of equation (36). To see this, note that

$$\frac{1}{n-m}\sum_{s=1}^{n-m} f_{t+m}^{(s)} = y_{t+m}^{(n-m)}, \qquad \frac{1}{n-m}\sum_{s=1}^{n-m} f_t^{(s+m)} = y_t^{(n)} + \frac{m}{n-m}\left(y_t^{(n)} - y_t^{(m)}\right), \tag{41}$$

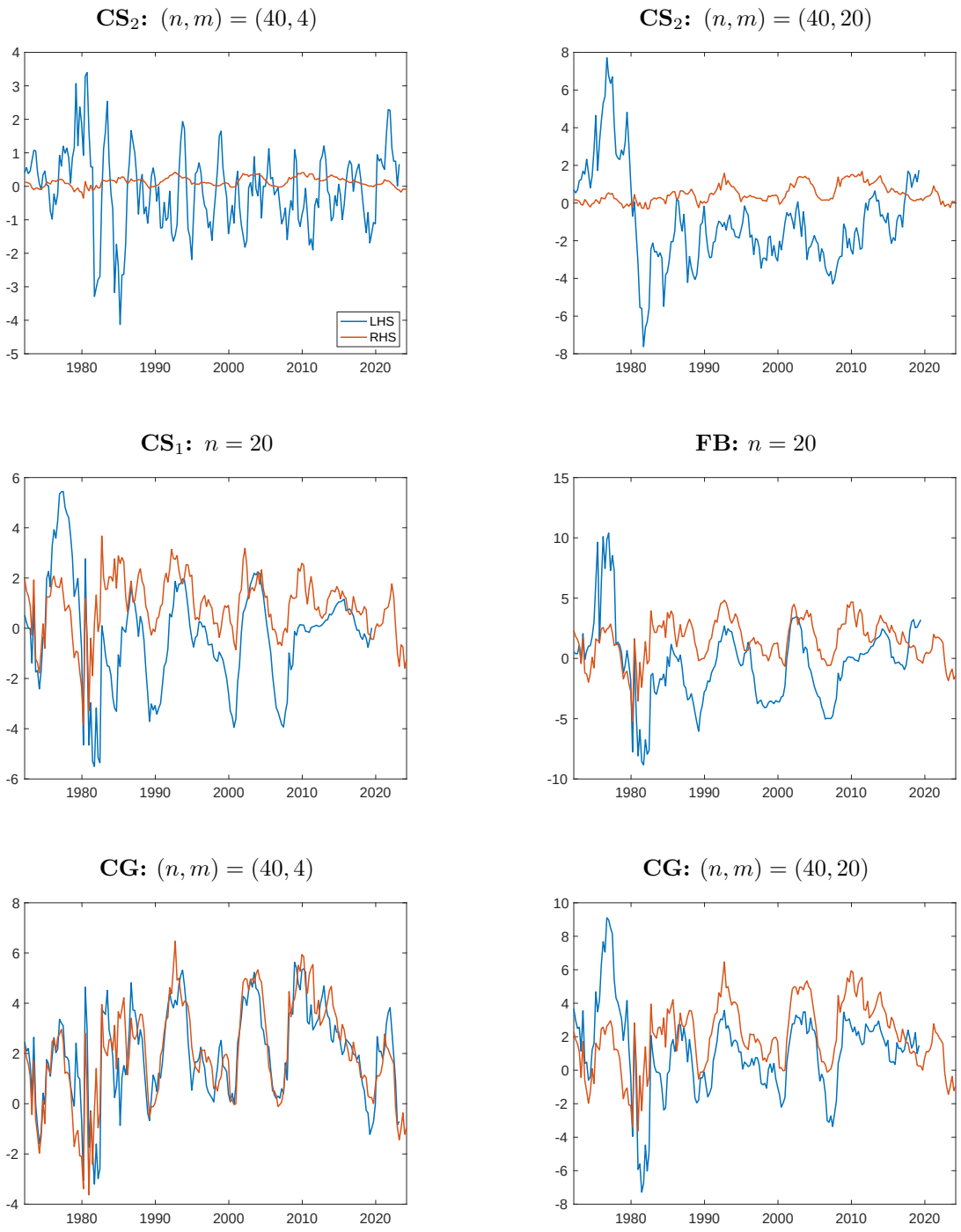
and $v_{\mathrm{CS}_2,t+n-1}^{(n)} = \frac{1}{n-m}\sum_{s=1}^{n-m} \epsilon_{t+m}^{(s,m)}$.

Although FB, CS_1, CS_2 can all be derived from CG, their statistical properties may be very different. To illustrate this, we use quarterly bond yield data from Gurkaynak et al. (2007) for maturities up to ten years for the sample period 1972Q1 to 2024Q1. Figure 4 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.[24] When $m = 20$, the point estimate of β rises but is close to zero for most maturities. The results of 1,000 bootstrap draws are presented in grey.[25] 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

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

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

Figure 5. 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:Q1–2024:Q1.



evidence against the expectations hypothesis for the CS_2 test is generally weaker than the point estimate would suggest as most bootstrap draws lie above it.

Figure 5 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 5 shows that when $m = 4$, the dependent and explanatory variables have strikingly different persistence and volatility properties. This unbalancedness comes about because the dependent variable in the CS_2 specification involves $y_{t+m}^{(n-m)} - y_t^{(n)}$ rather than just $y_{t+m}^{(n-m)}$ as in equation (41). In fact, the OLS estimate is no longer negative if we regress $y_{t+m}^{(n-m)} - y_t^{(1)}$ on $y_t^{(n)} - y_t^{(1)} + \frac{m}{n-m}(y_t^{(n)} - y_t^{(m)})$. 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 4 provides a translation of the statistical problems induced by the unbalanced specification – as hinted by the time-series plots in Figure 5 – which is mapped to the sampling variability of the OLS estimator.

Conversely, the CS_1 and FB tests, in the middle row of Figure 4, 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 (36) directly. In the bottom row of Figure 4, we present results for $m = 4$ and $m = 20$. 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 the bottom row of Figure 5.

This application highlights the flexibility of our bootstrap procedure and how it can offer insights to existing results in the literature. In particular, we have shown how different the sampling properties of the OLS estimator can be for seemingly innocuous choices of how a regression-based test of the EH is implemented.

4.3 Estimation and Inference in Probability of Recession Models

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).[26] 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) available toward the end of the subsequent quarter, rather than the current, revised series. Our z_t then becomes $z_t = \left\{ f_{t+1}^{(N)}, dr_{t+1}^{(2)}, \ldots, dr_{t+1}^{(N)}, G_{t+1} \right\}$ and we can easily block bootstrap this augmented matrix as detailed in Section 2.

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

$$P\left(H_{t+h} = 1 \middle| x_t\right) = \Phi\left(x_t'\beta\right) \tag{42}$$

for some $h \geq 1$, where H_t is a transformation of G_t and x_t are term-structure variables defined below. We work with two choices for H_t: first, we define $H_{1,t+h} = \mathbb{1}\{G_{t+h} \leq 0\}$; second, we define $H_{2,t+h} = \mathbb{1}\left\{\bigcup_{r=t+1}^{t+h} \{G_r \leq 0\}\right\}$. 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 (42). 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

$$\hat{\beta}^{\text{bc}} = \hat{\beta} - \frac{1}{B}\sum\nolimits_{b=1}^{B} \hat{d}_b^*. \tag{43}$$

[26] 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.

We can also construct bootstrap-based confidence intervals as

$$C_\alpha(\beta) = \left[\hat{\beta} - \hat{d}^*_{(1-\alpha)}, \hat{\beta} - \hat{d}^*_{(\alpha)}\right], \tag{44}$$

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 a constant and 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, which is not explicitly imposed in our bootstrap data, we use the deviation of the 3-month yield from its 3-year moving average.[27] 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 order to robustify the results, we use a leave-one-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 6 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 1971Q3–2024Q1. The top chart shows the time series of the fitted conditional probability that $H_{2,t} = 1$ 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 estimated probability, as of the beginning of 2024, is over 80% 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, relative to 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).

In the top two charts, we also report the fitted values without bias correction (red lines). We observe that for small fitted probabilities – periods associated with a wider term spread – the bias-

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

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 illustrates the advantages of our bootstrap approach.

The bottom two charts present pointwise 68% boostrap-based confidence intervals for the probability of recession for each specification. We observe that the bootstrap-based uncertainty measure around the estimated probability tends to be wider when the estimated probability of recession is at intermediate values. In the SA, we report the results for $H_{1,t}$ which are largely similar. One important difference is that the width of the bootstrap-based confidence intervals is shorter for $H_{1,t}$ than for $H_{2,t}$. This is an appealing result of our inference approach as it is consistent with the intuition that $H_{2,t}$ represents a more serially correlated process than $H_{1,t}$, 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. 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 method stands in sharp contrast to the conventional approach of committing to a factor dimension and specific parametric form for the cross-sectional dependence and time-series persistence.

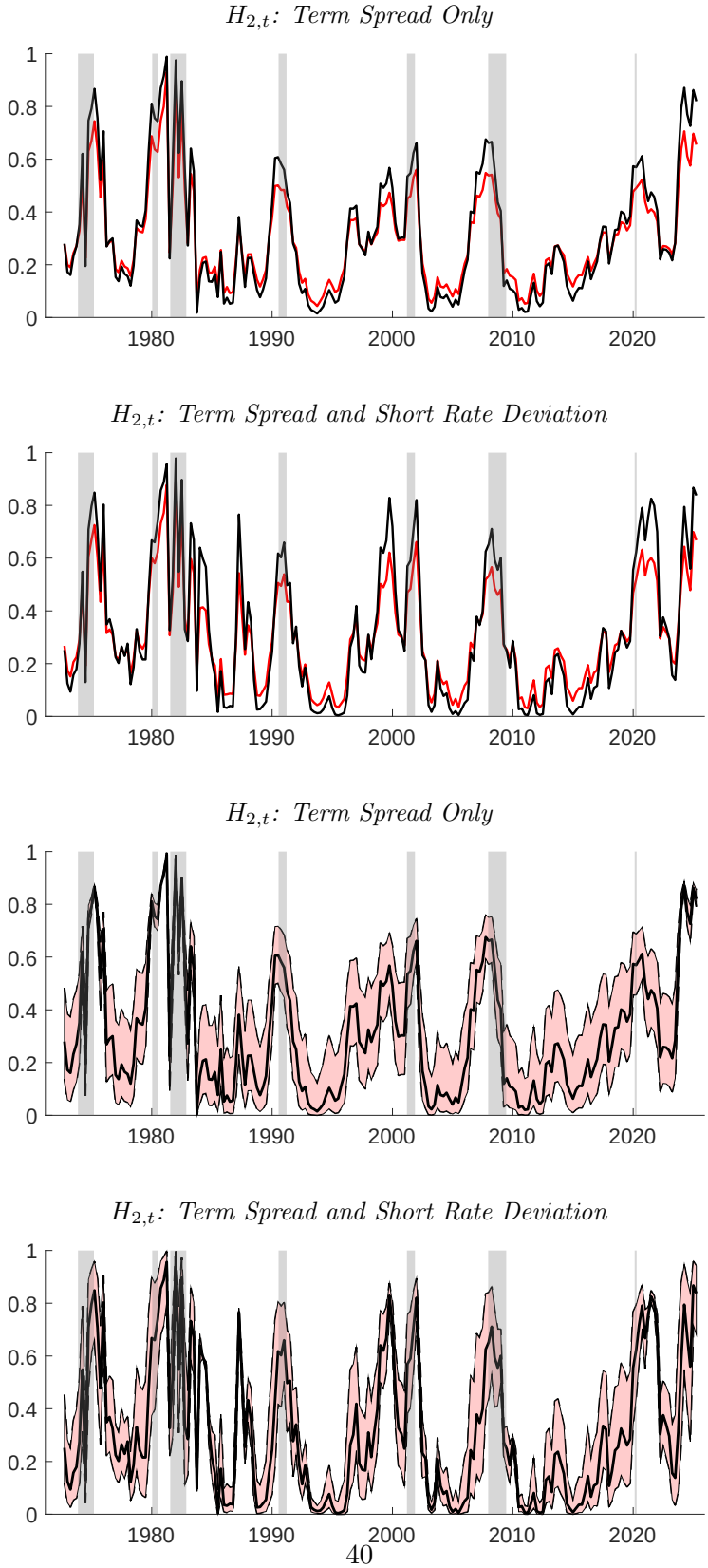
In our main empirical application, we revisit the predictive power of slow-moving economic fundamentals – trend inflation and the equilibrium real rate – for future bond returns (Cieslak and Povala 2015; Bauer and Rudebusch 2020). We first clarify why existing methods used to conduct inference in this setting, such as HAC/HAR variance estimators or the bootstrap method of Bauer and Hamilton (2018), are unreliable. As in Cieslak and Povala (2015), we find that trend inflation significantly forecasts future bond returns beyond the information contained in the current yield curve. In contrast, we cannot find any evidence to suggest that the equilibrium real rate has predictive content for future bond returns. We also show how our bootstrap can shed light on the

sampling properties of the OLS estimator in regression-based tests of the expectations hypothesis. 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.

The proposed resampling scheme can be further 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. Our bootstrap method can also be adapted to 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.

Figure 6. Probability of Recession

This figure plots the fitted values of $P(H_{2,t} = 1|\, x_t)$ 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 (red) line represents the estimated probability with (without) the bootstrap bias correction; pink shading denotes 68% pointwise confidence intervals for the probability of recession. Grey shading denotes NBER recessions. The sample period is 1971Q3–2024Q1.



Appendix

A Data Sources and Additional Details

The primary data set we use is continuously-compounded, zero-coupon Treasury yields from Gurkaynak et al. (2007) (hereafter, "GSW").[28] We use both monthly and quarterly data up to a maximum of a fifteen-year maturity. In previous versions of this paper, we also utilized the yield curve data from Liu and Wu (2021) and our results are robust to this different data set.

Section 2 In Figure 1, we use end-of-quarter GSW yields for quarterly maturities up to 10 years for the sample period from 1972Q1–2022Q4. The remaining data sources for this figure are as follows. First, the data on oil futures consist 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 are from Constantinides et al. (2013), and aggregated and sorted by maturity as in He et al. (2017).[29] 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.[30]

Section 3 To calibrate the simulations, we use end-of-month observations of yields for annual maturities ranging from one to ten years, based on the GSW data. To calibrate the external predictors, we use monthly industrial production (Fred mnemonic: `INDPRO`) and the core CPI price index (Fred mnemonic: `CPILFESL`). All necessary parameters utilized in the simulation experiments are estimated over the sample 1972M1–2022M12.

Section 4.1 For Tables 7 and 8, we use the same data sources as in BR. Our bond data are based on end-of-quarter yields for 3-month and 6-month maturities from the H.15 release[31] and end-of-quarter yields for annual maturities between one year and 15 years from Gurkaynak et al. (2007). We use data for the PTR and year-over-year core PCE inflation directly from the replication code of Bauer and Rudebusch (2020) (hereafter, "BR").[32] For Table 9, we use the core CPI index as in Section 3. We construct quarterly observations by averaging the index level within the quarter. For the results in Table 9, we construct r_t^* using the end-of-quarter observation for the 3-month yield rather than the average of the last month of the

[28] Data are available at `https://www.federalreserve.gov/pubs/feds/2006/200628/200628abs.html`.
[29] We thank the authors for making the data available.
[30] `https://data.giss.nasa.gov/gistemp/graphs/`
[31] Fred mnemonics: `DTB3` and `DTB6`, respectively. More information is available at `https://www.federalreserve.gov/releases/h15/`.
[32] Replication code is available at `https://www.openicpsr.org/openicpsr/project/115622/version/V1/view`.

quarter as done in BR. It is important to emphasize that using the end-of-quarter 3-month yield shifts r_t^* by minimal amounts. For example, over the 1971Q4–2018Q1 sample, the average absolute difference between the two series is 2.8 basis points and the maximum absolute difference is 5.6 basis points.

For both the CG bootstrap and the BH bootstrap, we consider different bootstrap samples for Specifications (I), (II), and (III) separately. For example, when conducting bootstrap inference for Specification (I), we treat π_t^* as the only external predictor.

For the CG bootstrap, we want to mimic the empirical implementation in BR and Cieslak and Povala (2015) and so we require a ten-year burn-in period. To do so, we pre-whiten using a VAR(1) specification on the sample 1971Q4–2018Q1 which is of length T but generate a bootstrapped sample for $T+40$ observations. We use initial conditions for the forward curve, year-over-year core PCE inflation, and year-over-year core CPI inflation as the realized data in 1961Q4. The PTR series is not available for 1961Q4 and so we instead use the consensus forecast of annualized 6-month inflation, 6-months in the future, from the December 1961 Livingston Survey, as the initial condition.[33] For the full-sample (1971Q4–2018Q1) regression results, we drop the first 40 observations from the bootstrapped data to construct bootstrap-based OLS estimates, Eicker-White standard errors ($h = 1$), and Newey-West standard errors ($h = 6$).[34] For the subsample starting in 1985Q1, we drop the first 93 observations from the bootstrapped data to construct OLS and standard error estimates. Within the CG bootstrap, we construct r_t^* using the end-of-quarter 3-month yield less the year-over-year inflation rate. This approximates the implementation in the realized data in Tables 7 and 8 but is exact for Table 9. However, it is important to emphasize that this has negligible effects on our inference results and is instead done only for convenience by avoiding the need to additionally bootstrap the monthly average of the last month of the quarter for the 3-month yield.

Section 4.3 We use end-of-quarter GSW yields for quarterly maturities up to 10 years. We use the third release of GDP (often referred to as the "final" release) available toward the end of the subsequent quarter.[35] The sample period is 1971Q3–2024Q1.

[33] Data are available here: https://www.philadelphiafed.org/surveys-and-data/real-time-data-research/livingston-survey.

[34] Although BR do not report results for $h = 4$, estimation and inference results for this forecast horizon are available in their replication code. We follow BR and use the Newey-West estimator with 6 lags in our implementation of the CG bootstrap. Note that all of the results and conclusions in Section 4 are robust to using 4 lags when implementing the CG bootstrap instead.

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

The replication code and data are available in the Harvard Dataverse at https://doi.org/10.7910/DVN/DOLUXM.

References

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

Audoly, R., Almuzara, M., Crump, R., Melcangi, D., Xing, R., 2023. How large are inflation revisions? the difficulty of monitoring prices in real time. Liberty Street Economics Blog.

Bauer, M., Chernov, M., 2024. Interest rate skewness and biased beliefs. Journal of Finance 79, 173–217.

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.

Bianchi, D., Büchner, M., Tamoni, A., 2020. Bond risk premiums with machine learning. Review of Financial Studies 34, 1046–1089.

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., Gospodinov, N., 2022. On the factor structure of bond returns. Econometrica 90, 295–314.

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.

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

Filipović, D., Pelger, M., Ye, Y., 2024. Shrinking the term structure. Working Paper No. 32472, NBER.

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.

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.

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

Kilian, L., Lütkepohl, H., 2017. Structural Vector Autoregressive Analysis. Cambridge University Press, Cambridge, U.K.

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.

Liu, Y., Wu, J. C., 2021. Reconstructing the yield curve. Journal of Financial Economics 3, 1395–1425.

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

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.

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

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., 2023. Pricing the bund term structure with linear regressions - without an observable short rate. Discussion Paper No 08/2023, Deutsche Bundesbank.

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.

The Wall Street Journal, 2000. Major price index is revised upward as a result of error.

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.

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.

Supplemental Appendix: "Deconstructing the Yield Curve"

Richard K. Crump
(New York Fed)

Nikolay Gospodinov
(Atlanta Fed)

Contents

SA-1 Bootstrap Validity **2**

SA-1.1 Assumptions 2

SA-1.2 Technical Results and Proofs 3

SA-2 Additional Simulation Results **6**

SA-2.1 No Bias Correction 6

SA-2.2 Additional Designs 10

SA-2.3 Results for 5% Nominal Size 13

SA-2.4 Inference Based on Asymptotic Approximations 17

SA-3 Probability of Recession: Additional Results **20**

SA-1 Bootstrap Validity

SA-1.1 Assumptions

In this appendix, we will write z_t from the main text as z_{nt} for full clarity. We first provide a definition of $\{z_{nt}\}$ as a mixing process. Let $F_{-\infty}^{n,t} = \sigma(..., z_{n(t-1)}, z_{nt})$ 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. Let Δ be a fixed, positive constant.

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 fixed;

(A1.e) the block size M diverges with T ($M_T \to \infty$) 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 forms of 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 grows 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 $B \equiv \lim_{T\to\infty}\text{Var}\left(T^{-1/2}\sum_{t=1}^{T-h}(x_t - \bar{x})\varepsilon_{t+h}^{(n)}\right)$ is positive definite with $\det(B)$ bounded away from zero.

Assumption (A2.a) imposes stationarity on the predictors and the errors. Stationary but highly persistent predictors, such as the yields and forward rates, 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) $\int_{-\infty}^{\infty}|\omega(x)|\text{d}x < \infty$, (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).

SA-1.2 Technical Results and Proofs

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 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.

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 $\varepsilon^{(n)}_{t+h}$ and x_t. This is done in Assumption A2. Furthermore, Assumption A3 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 $(x_t - \bar{x})\hat{\varepsilon}^{(n)}_{t+h}$. 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 $\hat{\varepsilon}^{(n)*}_{t+h}$ in these expressions. With this notation, define the t-statistic of the ith predictor $\hat{t}_i = (\hat{\beta}^{(n)}_i - \beta^{(n)}_i)/\text{se}(\hat{\beta}^{(n)}_i)$, where $\text{se}(\hat{\beta}^{(n)}) = \sqrt{\text{diag}(\hat{\Omega}_T)}$, and $t^*_i = (\hat{\beta}^{(n)*}_i - \hat{\beta}^{(n)}_i)/\text{se}(\hat{\beta}^*_i)$, where $\text{se}(\hat{\beta}^{(n)*}) = \sqrt{\text{diag}(\hat{\Omega}^*_T)}$. The next theorem establishes the validity of the bootstrap method.

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

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

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

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

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

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

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

$$\hat{A}_T^{-1} - A_T^{-1} \xrightarrow{P} 0, \qquad \hat{A}_T^{*-1} - \hat{A}_T^{-1} \xrightarrow{P^*} 0, \qquad \hat{B}^*_T - B_T \xrightarrow{P^*} 0,$$

in probability, and

$$B_T^{-1/2}\frac{1}{\sqrt{T}}\sum_{t=1}^{T-h}(x^*_t - \bar{x}^*)\varepsilon^{(n)*}_{t+h} \xrightarrow{d^*} N(0, I),$$

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

Proof of Theorem 1: 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

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

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

SA-2 Additional Simulation Results

SA-2.1 No Bias Correction

To assess the effect of the bias-corrected VAR(1) estimates on our procedure, we report the counterpart of Tables 1–6 for both the CG bootstrap and the BH bootstrap using unadjusted VAR(1) estimates. In Tables SA.1, SA.3, and SA.5, we report the empirical size for the three different simulation designs considered in the main text. The results show that the CG bootstrap with and without bias correction produces very similar results. This is also the case for the empirical power which is shown in Tables SA.2, SA.4, and SA.6.

Table SA.1. Size: Three-Factor VAR(1) [No Bias Correction]. This table presents empirical size for the nonparametric bootstrap introduced in Section 2.2 (CG) and the parametric bootstrap of Bauer and Hamilton (2018) (BH). Both procedures use OLS estimates rather than the bias-correction approach discussed 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 $(g_{1t}, g_{2t}, g_{3t}) = (y_t^{(12)}, f_t^{(60,12)}, f_t^{(120,12)})$. Each row reports results for bond returns of the corresponding maturity. Based on 5,000 simulations and 399 bootstrap replications per simulation.

	CG Bootstrap			**BH Bootstrap**		
	g_{1t}	g_{2t}	g_{3t}	g_{1t}	g_{2t}	g_{3t}
2y	0.108	0.110	0.100	0.068	0.024	0.013
3y	0.101	0.113	0.102	0.067	0.019	0.004
4y	0.099	0.122	0.105	0.065	0.013	0.000
5y	0.100	0.120	0.109	0.063	0.013	0.000
6y	0.101	0.118	0.107	0.059	0.010	0.000
7y	0.102	0.122	0.105	0.054	0.008	0.000
8y	0.104	0.122	0.108	0.053	0.008	0.000
9y	0.107	0.123	0.107	0.053	0.011	0.000
10y	0.107	0.123	0.108	0.050	0.008	0.000

Table SA.2. Power: Three-Factor VAR(1) [No Bias Correction]. This table presents empirical size for the nonparametric bootstrap introduced in Section 2.2 (CG) and the parametric bootstrap of Bauer and Hamilton (2018) (BH). Both procedures use OLS estimates rather than the bias-correction approach discussed 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 $(g_{1t}, g_{2t}, g_{3t}) = (y_t^{(12)}, f_t^{(60,12)}, f_t^{(120,12)})$. Each row reports results for bond returns of the corresponding maturity. Based on 5,000 simulations and 399 bootstrap replications per simulation.

	CG Bootstrap			**BH Bootstrap**		
	g_{1t}	g_{2t}	g_{3t}	g_{1t}	g_{2t}	g_{3t}
2y	0.257	0.247	0.101	0.264	0.160	0.004
3y	0.422	0.380	0.105	0.414	0.257	0.007
4y	0.538	0.446	0.114	0.523	0.310	0.005
5y	0.616	0.461	0.116	0.593	0.332	0.002
6y	0.659	0.440	0.110	0.630	0.307	0.000
7y	0.681	0.390	0.106	0.648	0.261	0.000
8y	0.688	0.333	0.116	0.650	0.203	0.000
9y	0.686	0.276	0.146	0.640	0.164	0.000
10y	0.676	0.220	0.198	0.631	0.086	0.000

Table SA.3. Size: Five-Factor VAR(1) [No Bias Correction]. This table presents empirical size for the nonparametric bootstrap introduced in Section 2.2 (CG) and the parametric bootstrap of Bauer and Hamilton (2018) (BH). Both procedures use OLS estimates rather than the bias-correction approach discussed 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 $(g_{1t}, g_{2t}, g_{3t}, g_{4t}, g_{5t}) = (y_t^{(12)}, f_t^{(36,12)}, f_t^{(60,12)}, f_t^{(84,12)}, f_t^{(120,12)})$. Each row reports results for bond returns of the corresponding maturity. Based on 5,000 simulations and 399 bootstrap replications per simulation.

	CG Bootstrap					**BH Bootstrap**				
	g_{1t}	g_{2t}	g_{3t}	g_{3t}	g_{3t}	g_{1t}	g_{2t}	g_{3t}	g_{3t}	g_{3t}
2y	0.108	0.104	0.095	0.085	0.090	0.048	0.022	0.012	0.006	0.013
3y	0.101	0.101	0.096	0.087	0.089	0.035	0.010	0.012	0.007	0.012
4y	0.100	0.097	0.094	0.090	0.089	0.023	0.005	0.011	0.007	0.016
5y	0.096	0.096	0.094	0.091	0.091	0.019	0.005	0.011	0.010	0.015
6y	0.098	0.097	0.094	0.092	0.092	0.016	0.003	0.014	0.008	0.012
7y	0.098	0.095	0.094	0.092	0.090	0.016	0.003	0.012	0.007	0.010
8y	0.099	0.096	0.097	0.093	0.094	0.015	0.005	0.015	0.007	0.010
9y	0.098	0.100	0.096	0.095	0.094	0.016	0.006	0.015	0.008	0.013
10y	0.100	0.100	0.097	0.097	0.098	0.014	0.007	0.015	0.007	0.012

Table SA.4. Size: Five-Factor VAR(1) [No Bias Correction]. This table presents empirical size for the nonparametric bootstrap introduced in Section 2.2 (CG) and the parametric bootstrap of Bauer and Hamilton (2018) (BH). Both procedures use OLS estimates rather than the bias-correction approach discussed 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 $(g_{1t}, g_{2t}, g_{3t}, g_{4t}, g_{5t}) = (y_t^{(12)}, f_t^{(36,12)}, f_t^{(60,12)}, f_t^{(84,12)}, f_t^{(120,12)})$. Each row reports results for bond returns of the corresponding maturity. Based on 5,000 simulations and 399 bootstrap replications per simulation.

	CG Bootstrap					**BH Bootstrap**				
	g_{1t}	g_{2t}	g_{3t}	g_{3t}	g_{3t}	g_{1t}	g_{2t}	g_{3t}	g_{3t}	g_{3t}
2y	0.356	0.506	0.545	0.569	0.475	0.020	0.000	0.000	0.000	0.000
3y	0.396	0.431	0.462	0.523	0.479	0.070	0.000	0.000	0.000	0.000
4y	0.397	0.356	0.400	0.499	0.490	0.105	0.000	0.000	0.000	0.000
5y	0.392	0.310	0.371	0.503	0.514	0.107	0.000	0.000	0.000	0.000
6y	0.395	0.289	0.368	0.527	0.541	0.136	0.000	0.000	0.000	0.000
7y	0.400	0.280	0.386	0.561	0.556	0.135	0.000	0.000	0.000	0.000
8y	0.409	0.280	0.408	0.578	0.559	0.136	0.000	0.000	0.000	0.000
9y	0.418	0.282	0.425	0.587	0.542	0.126	0.000	0.000	0.000	0.000
10y	0.420	0.282	0.432	0.579	0.509	0.128	0.000	0.000	0.000	0.000

Table SA.5. Size: Three-Factor VAR(1) with External Predictors [No Bias Correction]. This table presents empirical size for the nonparametric bootstrap introduced in Section 2.2 (CG) and the parametric bootstrap of Bauer and Hamilton (2018) (BH). Both procedures use OLS estimates rather than the bias-correction approach discussed 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 $(g_{1t}, g_{2t}, g_{3t}) = (y_t^{(12)}, f_t^{(60,12)}, f_t^{(120,12)})$ and the external predictors w_{1t} and w_{2t}. Each row reports results for bond returns of the corresponding maturity. Based on 5,000 simulations and 399 bootstrap replications per simulation.

	CG Bootstrap					**BH Bootstrap**				
	g_{1t}	g_{2t}	g_{3t}	w_{1t}	w_{2t}	g_{1t}	g_{2t}	g_{3t}	w_{1t}	w_{2t}
2y	0.111	0.110	0.095	0.104	0.104	0.067	0.021	0.013	0.112	0.106
3y	0.105	0.112	0.101	0.107	0.107	0.065	0.018	0.005	0.111	0.108
4y	0.101	0.115	0.100	0.108	0.106	0.062	0.012	0.001	0.111	0.107
5y	0.098	0.117	0.101	0.111	0.108	0.059	0.010	0.001	0.111	0.108
6y	0.096	0.113	0.101	0.109	0.108	0.055	0.008	0.000	0.107	0.108
7y	0.097	0.113	0.104	0.109	0.108	0.054	0.008	0.000	0.106	0.105
8y	0.095	0.114	0.100	0.109	0.104	0.051	0.009	0.000	0.107	0.106
9y	0.096	0.115	0.103	0.109	0.106	0.049	0.013	0.000	0.109	0.105
10y	0.099	0.115	0.104	0.109	0.104	0.046	0.008	0.000	0.109	0.106

Table SA.6. Power: Three-Factor VAR(1) with External Predictors [No Bias Correction]. This table presents empirical power for the nonparametric bootstrap introduced in Section 2.2 (CG) and the parametric bootstrap of Bauer and Hamilton (2018) (BH). Both procedures use OLS estimates rather than the bias-correction approach discussed 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 $(g_{1t}, g_{2t}, g_{3t}) = (y_t^{(12)}, f_t^{(60,12)}, f_t^{(120,12)})$. Each row reports results for bond returns of the corresponding maturity. Based on 5,000 simulations and 399 bootstrap replications per simulation.

	CG Bootstrap			**BH Bootstrap**		
	g_{1t}	g_{2t}	g_{3t}	g_{1t}	g_{2t}	g_{3t}
2y	0.222	0.245	0.099	0.238	0.166	0.004
3y	0.395	0.389	0.103	0.393	0.264	0.010
4y	0.514	0.458	0.111	0.509	0.318	0.006
5y	0.595	0.472	0.111	0.578	0.342	0.003
6y	0.637	0.446	0.105	0.615	0.316	0.001
7y	0.661	0.397	0.099	0.634	0.269	0.000
8y	0.669	0.341	0.113	0.638	0.212	0.000
9y	0.666	0.277	0.142	0.631	0.165	0.000
10y	0.659	0.227	0.191	0.622	0.089	0.000

SA-2.2 Additional Designs

Tables SA.7 and SA.8 report the empirical size and power for the five-factor model with a larger sample size than reported in the main text (Tables 3 and 4 in the main text, respectively). We use $T = 2000$ which shows that empirical size continues to be controlled well using our rule-of-thumb choice for M (see Table SA.7). Table SA.8 shows that the power of the CG bootstrap approaches one as the sample size increases. Moreover, the discrepancy between the power of the CG bootstrap and the O bootstrap is also reduced with the larger sample size. Note that even with the larger sample size, the misspecification of the factor space in the BH bootstrap implies that the finite-sample properties of this procedure remain poor.

Tables SA.9 and SA.10 report empirical size and power for the VAR(2) specification discussed in Section 3 of the main text. Although the CG bootstrap has relatively low power for inference on the coefficients associated with the lagged factors, this is also the case for the O bootstrap. Again, this is due to the relative proximity of these coefficients, that are calibrated to actual data, to zero. Finally, note that the misspecification of the time-series dynamics leads to poor finite-sample performance for the BH bootstrap under both the null and alternative hypotheses.

Table SA.7. Size: Five-Factor VAR(1) [T = 2,000]. This table presents empirical size for the three bootstrap methods: the nonparametric bootstrap introduced in Section 2.2 (CG); the parametric bootstrap of Bauer and Hamilton (2018) (BH); and the oracle bootstrap (O). The nominal level is 10% and the sample size is $T = 2,000$. Each column reports results for the t-test associated with the regressor $(g_{1t}, g_{2t}, g_{3t}, g_{4t}, g_{5t}) = (y_t^{(12)}, f_t^{(36,12)}, f_t^{(60,12)}, f_t^{(84,12)}, f_t^{(120,12)})$. Each row reports results for bond returns of the corresponding maturity. Based on 5,000 simulations and 399 bootstrap replications per simulation.

	CG Bootstrap					**BH Bootstrap**					**O Bootstrap**				
	g_{1t}	g_{2t}	g_{3t}	g_{4t}	g_{5t}	g_{1t}	g_{2t}	g_{3t}	g_{4t}	g_{5t}	g_{1t}	g_{2t}	g_{3t}	g_{4t}	g_{5t}
2y	0.090	0.092	0.086	0.080	0.094	0.032	0.002	0.000	0.000	0.000	0.098	0.096	0.101	0.107	0.111
3y	0.095	0.094	0.087	0.078	0.089	0.036	0.005	0.001	0.000	0.000	0.101	0.096	0.098	0.104	0.108
4y	0.094	0.094	0.084	0.079	0.089	0.031	0.013	0.001	0.001	0.000	0.100	0.095	0.102	0.107	0.108
5y	0.096	0.099	0.083	0.077	0.087	0.029	0.014	0.001	0.001	0.000	0.100	0.098	0.099	0.105	0.107
6y	0.097	0.099	0.085	0.077	0.087	0.025	0.015	0.001	0.000	0.000	0.103	0.098	0.097	0.103	0.106
7y	0.099	0.099	0.085	0.079	0.085	0.024	0.016	0.001	0.000	0.000	0.104	0.098	0.097	0.101	0.107
8y	0.102	0.099	0.086	0.079	0.083	0.025	0.014	0.001	0.000	0.000	0.103	0.095	0.097	0.102	0.105
9y	0.101	0.099	0.086	0.081	0.085	0.028	0.011	0.001	0.000	0.000	0.103	0.094	0.098	0.102	0.104
10y	0.103	0.100	0.086	0.080	0.086	0.028	0.012	0.001	0.000	0.000	0.104	0.094	0.096	0.100	0.103

Table SA.8. Power: Five-Factor VAR(1) [T = 2,000]. This table presents empirical power for the three bootstrap methods: the nonparametric bootstrap introduced in Section 2.2 (CG); the parametric bootstrap of Bauer and Hamilton (2018) (BH); and the oracle bootstrap (O). The nominal level is 10% and the sample size is $T = 2,000$. Each column reports results for the t-test associated with the regressor $(g_{1t}, g_{2t}, g_{3t}, g_{4t}, g_{5t}) = (y_t^{(12)}, f_t^{(36,12)}, f_t^{(60,12)}, f_t^{(84,12)}, f_t^{(120,12)})$. Each row reports results for bond returns of the corresponding maturity. Based on 5,000 simulations and 399 bootstrap replications per simulation.

	CG Bootstrap					**BH Bootstrap**					**O Bootstrap**				
	g_{1t}	g_{2t}	g_{3t}	g_{4t}	g_{5t}	g_{1t}	g_{2t}	g_{3t}	g_{4t}	g_{5t}	g_{1t}	g_{2t}	g_{3t}	g_{4t}	g_{5t}
2y	0.904	0.932	0.977	0.986	0.962	0.193	0.000	0.000	0.000	0.000	0.920	0.945	0.979	0.990	0.967
3y	0.914	0.877	0.948	0.980	0.963	0.521	0.000	0.000	0.000	0.000	0.924	0.884	0.953	0.985	0.968
4y	0.900	0.792	0.910	0.970	0.968	0.638	0.000	0.000	0.000	0.000	0.914	0.790	0.916	0.979	0.974
5y	0.887	0.711	0.888	0.974	0.976	0.598	0.000	0.000	0.000	0.000	0.897	0.712	0.898	0.980	0.981
6y	0.880	0.670	0.886	0.979	0.983	0.654	0.000	0.000	0.000	0.000	0.887	0.670	0.899	0.985	0.987
7y	0.877	0.659	0.898	0.986	0.985	0.629	0.000	0.000	0.000	0.000	0.884	0.660	0.911	0.988	0.989
8y	0.877	0.664	0.914	0.988	0.984	0.602	0.000	0.000	0.000	0.000	0.882	0.663	0.922	0.989	0.988
9y	0.879	0.673	0.927	0.988	0.980	0.533	0.000	0.000	0.000	0.000	0.885	0.674	0.933	0.989	0.986
10y	0.879	0.676	0.931	0.986	0.973	0.533	0.000	0.000	0.000	0.000	0.887	0.678	0.935	0.987	0.977

Table SA.9. Size: Three-Factor VAR(2). This table presents empirical size for the three bootstrap methods: the nonparametric bootstrap introduced in Section 2.2 (CG); the parametric bootstrap of Bauer and Hamilton (2018) (BH); and the oracle bootstrap (O). The nominal level is 10% and the sample size is $T = 600$. Each column reports results for the t-test associated with the regressor $(g_{1t}, g_{2t}, g_{3t}) = (y_t^{(12)}, f_t^{(60,12)}, f_t^{(120,12)})$. Each row reports results for bond returns of the corresponding maturity. Based on 5,000 simulations and 399 bootstrap replications per simulation.

	CG Bootstrap						**BH Bootstrap**						**O Bootstrap**					
	g_{1t}	g_{2t}	g_{3t}	$g_{1(t-1)}$	$g_{2(t-1)}$	$g_{3(t-1)}$	g_{1t}	g_{2t}	g_{3t}	$g_{1(t-1)}$	$g_{2(t-1)}$	$g_{3(t-1)}$	g_{1t}	g_{2t}	g_{3t}	$g_{1(t-1)}$	$g_{2(t-1)}$	$g_{3(t-1)}$
2y	0.129	0.110	0.105	0.092	0.079	0.076	0.041	0.023	0.004	0.029	0.000	0.000	0.106	0.103	0.100	0.109	0.103	0.103
3y	0.127	0.119	0.108	0.094	0.080	0.079	0.045	0.018	0.002	0.031	0.000	0.000	0.104	0.101	0.098	0.110	0.102	0.107
4y	0.127	0.123	0.108	0.099	0.081	0.081	0.051	0.012	0.001	0.037	0.000	0.000	0.105	0.101	0.096	0.109	0.104	0.107
5y	0.125	0.121	0.106	0.100	0.084	0.080	0.039	0.015	0.000	0.027	0.000	0.000	0.106	0.101	0.099	0.108	0.106	0.107
6y	0.123	0.121	0.106	0.103	0.081	0.080	0.044	0.009	0.000	0.030	0.000	0.000	0.104	0.099	0.097	0.110	0.106	0.104
7y	0.122	0.121	0.105	0.105	0.083	0.079	0.042	0.007	0.000	0.028	0.000	0.000	0.099	0.099	0.097	0.110	0.104	0.102
8y	0.118	0.122	0.107	0.107	0.082	0.077	0.042	0.007	0.000	0.027	0.000	0.000	0.099	0.102	0.097	0.110	0.105	0.102
9y	0.116	0.122	0.107	0.108	0.084	0.075	0.041	0.009	0.000	0.026	0.000	0.000	0.098	0.100	0.097	0.110	0.105	0.100
10y	0.116	0.122	0.110	0.107	0.083	0.075	0.038	0.006	0.000	0.020	0.000	0.000	0.098	0.101	0.098	0.110	0.104	0.098

Table SA.10. Power: Three-Factor VAR(2). This table presents empirical power for the three bootstrap methods: the nonparametric bootstrap introduced in Section 2.2 (CG); the parametric bootstrap of Bauer and Hamilton (2018) (BH); and the oracle bootstrap (O). The nominal level is 10% and the sample size is $T = 600$. Each column reports results for the t-test associated with the regressor $(g_{1t}, g_{2t}, g_{3t}) = (y_t^{(12)}, f_t^{(60,12)}, f_t^{(120,12)})$. Each row reports results for bond returns of the corresponding maturity. Based on 5,000 simulations and 399 bootstrap replications per simulation.

	CG Bootstrap						**BH Bootstrap**						**O Bootstrap**					
	g_{1t}	g_{2t}	g_{3t}	$g_{1(t-1)}$	$g_{2(t-1)}$	$g_{3(t-1)}$	g_{1t}	g_{2t}	g_{3t}	$g_{1(t-1)}$	$g_{2(t-1)}$	$g_{3(t-1)}$	g_{1t}	g_{2t}	g_{3t}	$g_{1(t-1)}$	$g_{2(t-1)}$	$g_{3(t-1)}$
2y	0.238	0.228	0.102	0.121	0.085	0.076	0.087	0.140	0.003	0.000	0.000	0.000	0.225	0.218	0.097	0.125	0.105	0.108
3y	0.373	0.348	0.119	0.115	0.084	0.082	0.193	0.224	0.008	0.001	0.000	0.000	0.370	0.327	0.109	0.118	0.105	0.112
4y	0.485	0.416	0.130	0.109	0.085	0.081	0.259	0.271	0.004	0.001	0.000	0.000	0.483	0.389	0.115	0.109	0.106	0.112
5y	0.556	0.437	0.122	0.109	0.086	0.080	0.370	0.299	0.002	0.001	0.000	0.000	0.554	0.404	0.106	0.107	0.105	0.111
6y	0.599	0.415	0.106	0.104	0.085	0.076	0.385	0.270	0.000	0.001	0.000	0.000	0.600	0.383	0.099	0.101	0.106	0.109
7y	0.620	0.368	0.107	0.103	0.086	0.075	0.407	0.227	0.000	0.001	0.000	0.000	0.624	0.337	0.098	0.100	0.107	0.105
8y	0.627	0.310	0.129	0.099	0.089	0.074	0.413	0.168	0.000	0.000	0.000	0.000	0.630	0.282	0.115	0.101	0.104	0.102
9y	0.624	0.255	0.169	0.097	0.088	0.075	0.412	0.125	0.000	0.000	0.000	0.000	0.626	0.224	0.158	0.098	0.107	0.099
10y	0.611	0.200	0.228	0.095	0.089	0.076	0.404	0.053	0.000	0.000	0.000	0.000	0.613	0.175	0.219	0.098	0.105	0.098

SA-2.3 Results for 5% Nominal Size

Tables SA.11–SA.16 report empirical size and power using a nominal level of 5% rather than 10% presented in Tables 1–6 in the main text. The results for this alternative choice are qualitatively similar to what is described and summarized in the main text.

Table SA.11. Size: Three-Factor VAR(1) [5% Nominal Size]. This table presents empirical size for the three bootstrap methods: the nonparametric bootstrap introduced in Section 2.2 (CG); the parametric bootstrap of Bauer and Hamilton (2018) (BH); and the oracle bootstrap (O). The nominal level is 5% and the sample size is $T = 600$. Each column reports results for the t-test associated with the regressor $(g_{1t}, g_{2t}, g_{3t}) = (y_t^{(12)}, f_t^{(60,12)}, f_t^{(120,12)})$. Each row reports results for bond returns of the corresponding maturity. Based on 5,000 simulations and 399 bootstrap replications per simulation.

	CG Bootstrap			**BH Bootstrap**			**O Bootstrap**		
	g_{1t}	g_{2t}	g_{3t}	g_{1t}	g_{2t}	g_{3t}	g_{1t}	g_{2t}	g_{3t}
2y	0.061	0.049	0.040	0.020	0.007	0.004	0.050	0.047	0.047
3y	0.060	0.055	0.044	0.020	0.005	0.001	0.049	0.048	0.049
4y	0.062	0.059	0.046	0.020	0.003	0.000	0.051	0.047	0.049
5y	0.061	0.060	0.046	0.020	0.002	0.000	0.052	0.046	0.049
6y	0.060	0.062	0.048	0.020	0.001	0.000	0.051	0.048	0.049
7y	0.061	0.063	0.046	0.019	0.001	0.000	0.051	0.047	0.049
8y	0.059	0.062	0.044	0.018	0.001	0.000	0.049	0.045	0.050
9y	0.059	0.062	0.043	0.017	0.001	0.000	0.048	0.046	0.049
10y	0.059	0.061	0.044	0.016	0.001	0.000	0.049	0.047	0.047

Table SA.12. Power: Three-Factor VAR(1) [5% Nominal Size]. This table presents empirical power for the three bootstrap methods: the nonparametric bootstrap introduced in Section 2.2 (CG); the parametric bootstrap of Bauer and Hamilton (2018) (BH); and the oracle bootstrap (O). The nominal level is 5% and the sample size is $T = 600$. Each column reports results for the t-test associated with the regressor $(g_{1t}, g_{2t}, g_{3t}) = (y_t^{(12)}, f_t^{(60,12)}, f_t^{(120,12)})$. Each row reports results for bond returns of the corresponding maturity. Based on 5,000 simulations and 399 bootstrap replications per simulation.

	CG Bootstrap			**BH Bootstrap**			**O Bootstrap**		
	g_{1t}	g_{2t}	g_{3t}	g_{1t}	g_{2t}	g_{3t}	g_{1t}	g_{2t}	g_{3t}
2y	0.162	0.152	0.039	0.156	0.092	0.001	0.198	0.145	0.052
3y	0.300	0.264	0.047	0.289	0.164	0.002	0.348	0.245	0.050
4y	0.417	0.331	0.058	0.384	0.195	0.001	0.461	0.311	0.056
5y	0.498	0.342	0.057	0.444	0.212	0.000	0.534	0.318	0.055
6y	0.535	0.323	0.050	0.483	0.163	0.000	0.580	0.299	0.051
7y	0.563	0.287	0.044	0.502	0.123	0.000	0.602	0.255	0.051
8y	0.572	0.236	0.046	0.507	0.082	0.000	0.607	0.206	0.059
9y	0.573	0.186	0.062	0.505	0.061	0.000	0.603	0.161	0.079
10y	0.565	0.139	0.088	0.497	0.021	0.000	0.593	0.119	0.113

Table SA.13. Size: Five-Factor VAR(1) [5% Nominal Size]. This table presents empirical size for the three bootstrap methods: the nonparametric bootstrap introduced in Section 2.2 (CG); the parametric bootstrap of Bauer and Hamilton (2018) (BH); and the oracle bootstrap (O). The nominal level is 5% and the sample size is $T = 600$. Each column reports results for the t-test associated with the regressor $(g_{1t}, g_{2t}, g_{3t}, g_{4t}, g_{5t}) = (y_t^{(12)}, f_t^{(36,12)}, f_t^{(60,12)}, f_t^{(84,12)}, f_t^{(120,12)})$. Each row reports results for bond returns of the corresponding maturity. Based on 5,000 simulations and 399 bootstrap replications per simulation.

	CG Bootstrap					**BH Bootstrap**					**O Bootstrap**				
	g_{1t}	g_{2t}	g_{3t}	g_{4t}	g_{5t}	g_{1t}	g_{2t}	g_{3t}	g_{4t}	g_{5t}	g_{1t}	g_{2t}	g_{3t}	g_{4t}	g_{5t}
2y	0.059	0.050	0.050	0.047	0.053	0.019	0.007	0.008	0.003	0.007	0.048	0.057	0.051	0.051	0.050
3y	0.055	0.049	0.048	0.049	0.051	0.011	0.003	0.005	0.004	0.008	0.048	0.055	0.053	0.053	0.048
4y	0.053	0.050	0.048	0.052	0.051	0.007	0.001	0.003	0.004	0.009	0.050	0.057	0.052	0.051	0.048
5y	0.051	0.049	0.049	0.049	0.051	0.006	0.001	0.005	0.006	0.010	0.050	0.055	0.051	0.052	0.048
6y	0.051	0.048	0.049	0.048	0.050	0.005	0.000	0.006	0.006	0.008	0.050	0.052	0.051	0.052	0.046
7y	0.050	0.047	0.051	0.049	0.051	0.005	0.000	0.006	0.005	0.007	0.050	0.052	0.051	0.051	0.048
8y	0.050	0.045	0.050	0.050	0.053	0.005	0.001	0.006	0.004	0.008	0.052	0.053	0.052	0.051	0.047
9y	0.052	0.048	0.051	0.051	0.052	0.004	0.001	0.007	0.005	0.008	0.052	0.054	0.052	0.052	0.046
10y	0.053	0.049	0.051	0.052	0.053	0.004	0.001	0.008	0.005	0.009	0.052	0.053	0.053	0.049	0.046

Table SA.14. Power: Five-Factor VAR(1) [5% Nominal Size]. This table presents empirical power for the three bootstrap methods: the nonparametric bootstrap introduced in Section 2.2 (CG); the parametric bootstrap of Bauer and Hamilton (2018) (BH); and the oracle bootstrap (O). The nominal level is 5% and the sample size is $T = 600$. Each column reports results for the t-test associated with the regressor $(g_{1t}, g_{2t}, g_{3t}, g_{4t}, g_{5t}) = (y_t^{(12)}, f_t^{(36,12)}, f_t^{(60,12)}, f_t^{(84,12)}, f_t^{(120,12)})$. Each row reports results for bond returns of the corresponding maturity. Based on 5,000 simulations and 399 bootstrap replications per simulation.

	CG Bootstrap					**BH Bootstrap**					**O Bootstrap**				
	g_{1t}	g_{2t}	g_{3t}	g_{4t}	g_{5t}	g_{1t}	g_{2t}	g_{3t}	g_{4t}	g_{5t}	g_{1t}	g_{2t}	g_{3t}	g_{4t}	g_{5t}
2y	0.245	0.373	0.420	0.437	0.349	0.009	0.000	0.000	0.000	0.000	0.267	0.380	0.465	0.517	0.418
3y	0.279	0.318	0.336	0.386	0.353	0.034	0.000	0.000	0.000	0.000	0.300	0.322	0.392	0.472	0.422
4y	0.282	0.257	0.276	0.362	0.359	0.048	0.000	0.000	0.000	0.000	0.302	0.258	0.328	0.451	0.426
5y	0.280	0.215	0.252	0.364	0.384	0.049	0.000	0.000	0.000	0.000	0.297	0.219	0.307	0.459	0.452
6y	0.283	0.195	0.250	0.389	0.409	0.060	0.000	0.000	0.000	0.000	0.297	0.203	0.312	0.477	0.479
7y	0.287	0.190	0.270	0.414	0.427	0.060	0.000	0.000	0.000	0.000	0.300	0.199	0.326	0.498	0.498
8y	0.295	0.192	0.286	0.438	0.428	0.058	0.000	0.000	0.000	0.000	0.305	0.198	0.345	0.513	0.496
9y	0.304	0.194	0.301	0.446	0.411	0.050	0.000	0.000	0.000	0.000	0.310	0.202	0.353	0.518	0.479
10y	0.313	0.196	0.306	0.436	0.373	0.051	0.000	0.000	0.000	0.000	0.316	0.203	0.361	0.505	0.440

Table SA.15. Size: Three-Factor VAR(1) with External Predictors [5% Nominal Size]. This table presents empirical size for the three bootstrap methods: the nonparametric bootstrap introduced in Section 2.2 (CG); the parametric bootstrap of Bauer and Hamilton (2018) (BH); and the oracle bootstrap (O). The nominal level is 5% and the sample size is $T = 600$. Each column reports results for the t-test associated with the regressor $(g_{1t}, g_{2t}, g_{3t}) = (y_t^{(12)}, f_t^{(60,12)}, f_t^{(120,12)})$ and the external predictors w_{1t} and w_{2t}. Each row reports results for bond returns of the corresponding maturity. Based on 5,000 simulations and 399 bootstrap replications per simulation.

	CG Bootstrap					**BH Bootstrap**					**O Bootstrap**				
	g_{1t}	g_{2t}	g_{3t}	w_{1t}	w_{2t}	g_{1t}	g_{2t}	g_{3t}	w_{1t}	w_{2t}	g_{1t}	g_{2t}	g_{3t}	w_{1t}	w_{2t}
2y	0.068	0.051	0.039	0.050	0.052	0.021	0.006	0.003	0.049	0.054	0.052	0.050	0.048	0.050	0.052
3y	0.063	0.058	0.041	0.051	0.053	0.020	0.004	0.001	0.052	0.052	0.052	0.047	0.047	0.050	0.053
4y	0.063	0.058	0.043	0.053	0.052	0.021	0.002	0.000	0.053	0.055	0.051	0.049	0.049	0.051	0.052
5y	0.062	0.061	0.044	0.053	0.052	0.018	0.002	0.000	0.054	0.055	0.053	0.051	0.050	0.050	0.053
6y	0.061	0.063	0.043	0.055	0.051	0.018	0.002	0.000	0.051	0.053	0.051	0.051	0.049	0.048	0.053
7y	0.059	0.064	0.043	0.055	0.053	0.018	0.002	0.000	0.050	0.053	0.051	0.051	0.051	0.049	0.052
8y	0.058	0.065	0.041	0.058	0.053	0.018	0.001	0.000	0.048	0.054	0.051	0.052	0.050	0.050	0.053
9y	0.057	0.063	0.042	0.057	0.054	0.017	0.002	0.000	0.051	0.055	0.051	0.050	0.051	0.050	0.052
10y	0.055	0.062	0.042	0.057	0.055	0.016	0.001	0.000	0.051	0.054	0.050	0.050	0.051	0.051	0.053

Table SA.16. Power: Three-Factor VAR(1) with External Predictors [5% Nominal Size]. This table presents empirical power for the three bootstrap methods: the nonparametric bootstrap introduced in Section 2.2 (CG); the parametric bootstrap of Bauer and Hamilton (2018) (BH); and the oracle bootstrap (O). The nominal level is 5% and the sample size is $T = 600$. Each column reports results for the t-test associated with the regressor $(g_{1t}, g_{2t}, g_{3t}) = (y_t^{(12)}, f_t^{(60,12)}, f_t^{(120,12)})$. Each row reports results for bond returns of the corresponding maturity. Based on 5,000 simulations and 399 bootstrap replications per simulation.

	CG Bootstrap			**BH Bootstrap**			**O Bootstrap**		
	g_{1t}	g_{2t}	g_{3t}	g_{1t}	g_{2t}	g_{3t}	g_{1t}	g_{2t}	g_{3t}
2y	0.136	0.153	0.037	0.137	0.091	0.001	0.167	0.149	0.051
3y	0.274	0.269	0.045	0.264	0.170	0.003	0.314	0.251	0.050
4y	0.391	0.336	0.056	0.359	0.207	0.001	0.436	0.310	0.056
5y	0.472	0.349	0.055	0.432	0.221	0.001	0.515	0.324	0.055
6y	0.520	0.329	0.048	0.469	0.177	0.000	0.567	0.300	0.050
7y	0.548	0.287	0.041	0.491	0.133	0.000	0.591	0.257	0.051
8y	0.562	0.238	0.043	0.498	0.088	0.000	0.598	0.207	0.060
9y	0.563	0.186	0.056	0.494	0.067	0.000	0.595	0.159	0.078
10y	0.559	0.138	0.085	0.482	0.022	0.000	0.585	0.117	0.112

SA-2.4 Inference Based on Asymptotic Approximations

Here, we report the empirical size for procedures that rely on critical values obtained from asymptotic approximations to the t-statistic. We include results using the variance estimator of Newey and West (1987) with lag length $h = 12$ (labelled "NW") along with the two inference procedures discussed in Lazarus et al. (2018). The first approach uses the equal-weighted cosine (EWC) estimator of the long-run variance and a limiting t distribution (labelled "LLSW-EWC"). The second approach uses the variance estimator of Newey and West (1987) and a fixed-b asymptotic approximation (labelled "LLSW-NW"). We implement both of these tests exactly as described in Lazarus et al. (2018). Tables SA.17–SA.20 show that these procedures uniformly fail to control size in all of the simulation designs. The size distortions are particularly acute with extraneous external predictors as shown in Table SA.19. We 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.

Table SA.17. Size: Three-Factor VAR(1) [HAC/HAR Variance Estimator]. This table presents empirical size for three methods which rely on different sample variance estimators and asymptotic distributional approximations: Newey and West (1987) estimator with lag length of 12; the equal-weighted cosine estimator as in Lazarus et al. (2018); the Newey and West (1987) estimator as in Lazarus et al. (2018). The nominal level is 10% and the sample size is $T = 600$. Each column reports results for the t-test associated with the regressor $(g_{1t}, g_{2t}, g_{3t}) = (y_t^{(12)}, f_t^{(60,12)}, f_t^{(120,12)})$. Each row reports results for bond returns of the corresponding maturity. Based on 5,000 simulations and 399 bootstrap replications per simulation.

	NW			**LLSW-EWC**			**LLSW-NW**		
	g_{1t}	g_{2t}	g_{3t}	g_{1t}	g_{2t}	g_{3t}	g_{1t}	g_{2t}	g_{3t}
2y	0.263	0.202	0.200	0.194	0.147	0.146	0.208	0.157	0.155
3y	0.255	0.206	0.199	0.185	0.150	0.146	0.197	0.163	0.153
4y	0.242	0.208	0.196	0.182	0.153	0.145	0.190	0.161	0.149
5y	0.238	0.207	0.194	0.176	0.154	0.144	0.186	0.165	0.149
6y	0.236	0.212	0.194	0.178	0.153	0.141	0.187	0.167	0.152
7y	0.237	0.213	0.195	0.175	0.156	0.141	0.189	0.171	0.154
8y	0.235	0.213	0.193	0.177	0.160	0.140	0.186	0.171	0.155
9y	0.233	0.212	0.194	0.177	0.160	0.142	0.188	0.169	0.155
10y	0.232	0.212	0.194	0.174	0.157	0.142	0.187	0.166	0.155

Table SA.18. Size: Five-Factor VAR(1) [HAC/HAR Variance Estimators]. This table presents empirical size for three methods which rely on different sample variance estimators and asymptotic distributional approximations: Newey and West (1987) estimator with lag length of 12; the equal-weighted cosine estimator as in Lazarus et al. (2018); the Newey and West (1987) estimator as in Lazarus et al. (2018). The nominal level is 10% and the sample size is $T = 600$. Each column reports results for the t-test associated with the regressor $(g_{1t}, g_{2t}, g_{3t}, g_{4t}, g_{5t}) = (y_t^{(12)}, f_t^{(36,12)}, f_t^{(60,12)}, f_t^{(84,12)}, f_t^{(120,12)})$. Each row reports results for bond returns of the corresponding maturity. Based on 5,000 simulations and 399 bootstrap replications per simulation.

	NW					**LLSW-EWC**					**LLSW-NW**				
	g_{1t}	g_{2t}	g_{3t}	g_{4t}	g_{5t}	g_{1t}	g_{2t}	g_{3t}	g_{4t}	g_{5t}	g_{1t}	g_{2t}	g_{3t}	g_{4t}	g_{5t}
2y	0.209	0.177	0.179	0.177	0.180	0.157	0.141	0.142	0.140	0.138	0.169	0.147	0.148	0.147	0.146
3y	0.205	0.180	0.177	0.175	0.180	0.153	0.140	0.140	0.141	0.142	0.165	0.147	0.143	0.146	0.147
4y	0.202	0.181	0.174	0.176	0.181	0.149	0.140	0.139	0.142	0.144	0.162	0.148	0.146	0.147	0.147
5y	0.203	0.181	0.175	0.180	0.184	0.151	0.139	0.137	0.141	0.144	0.160	0.149	0.148	0.150	0.153
6y	0.202	0.179	0.174	0.180	0.188	0.149	0.139	0.139	0.142	0.145	0.162	0.148	0.148	0.149	0.154
7y	0.200	0.178	0.174	0.186	0.189	0.149	0.139	0.138	0.145	0.147	0.159	0.150	0.148	0.151	0.152
8y	0.197	0.180	0.179	0.183	0.191	0.147	0.138	0.139	0.145	0.143	0.158	0.147	0.148	0.154	0.152
9y	0.197	0.181	0.181	0.181	0.190	0.149	0.136	0.140	0.148	0.145	0.157	0.144	0.148	0.153	0.153
10y	0.196	0.180	0.182	0.180	0.190	0.150	0.136	0.143	0.148	0.148	0.157	0.145	0.149	0.153	0.153

Table SA.19. Size: Three-Factor VAR(1) with External Predictors [HAC/HAR Variance Estimators]. This table presents empirical size for three methods which rely on different sample variance estimators and asymptotic distributional approximations: Newey and West (1987) estimator with lag length of 12; the equal-weighted cosine estimator as in Lazarus et al. (2018); the Newey and West (1987) estimator as in Lazarus et al. (2018). The nominal level is 10% and the sample size is $T = 600$. Each column reports results for the t-test associated with the regressor $(g_{1t}, g_{2t}, g_{3t}) = (y_t^{(12)}, f_t^{(60,12)}, f_t^{(120,12)})$ and the external predictors w_{1t} and w_{2t}. Each row reports results for bond returns of the corresponding maturity. Based on 5,000 simulations and 399 bootstrap replications per simulation.

	NW					**LLSW-EWC**					**LLSW-NW**				
	g_{1t}	g_{2t}	g_{3t}	w_{1t}	w_{2t}	g_{1t}	g_{2t}	g_{3t}	w_{1t}	w_{2t}	g_{1t}	g_{2t}	g_{3t}	w_{1t}	w_{2t}
2y	0.277	0.212	0.210	0.255	0.182	0.212	0.155	0.157	0.189	0.142	0.228	0.166	0.165	0.198	0.148
3y	0.265	0.216	0.209	0.253	0.181	0.202	0.162	0.157	0.188	0.142	0.214	0.171	0.167	0.195	0.152
4y	0.257	0.221	0.209	0.251	0.179	0.195	0.162	0.155	0.186	0.143	0.207	0.171	0.165	0.192	0.152
5y	0.251	0.221	0.202	0.247	0.178	0.187	0.166	0.157	0.188	0.140	0.197	0.174	0.164	0.193	0.149
6y	0.249	0.222	0.202	0.251	0.179	0.185	0.167	0.154	0.187	0.139	0.196	0.176	0.166	0.193	0.149
7y	0.246	0.225	0.203	0.251	0.180	0.184	0.167	0.153	0.186	0.143	0.194	0.174	0.163	0.192	0.149
8y	0.244	0.224	0.200	0.251	0.182	0.182	0.166	0.150	0.185	0.141	0.193	0.178	0.160	0.192	0.148
9y	0.243	0.222	0.200	0.247	0.179	0.184	0.165	0.150	0.186	0.138	0.195	0.176	0.159	0.193	0.147
10y	0.243	0.222	0.197	0.246	0.177	0.182	0.164	0.150	0.183	0.140	0.193	0.174	0.160	0.192	0.147

Table SA.20. Size: Three-Factor VAR(2) [HAC/HAR Variance Estimators]. This table presents empirical size for three methods which rely on different sample variance estimators and asymptotic distributional approximations: Newey and West (1987) estimator with lag length of 12; the equal-weighted cosine estimator as in Lazarus et al. (2018); the Newey and West (1987) estimator as in Lazarus et al. (2018). The nominal level is 10% and the sample size is $T = 600$. Each column reports results for the t-test associated with the regressor $(g_{1t}, g_{2t}, g_{3t}) = (y_t^{(12)}, f_t^{(60,12)}, f_t^{(120,12)})$. Each row reports results for bond returns of the corresponding maturity. Based on 5,000 simulations and 399 bootstrap replications per simulation.

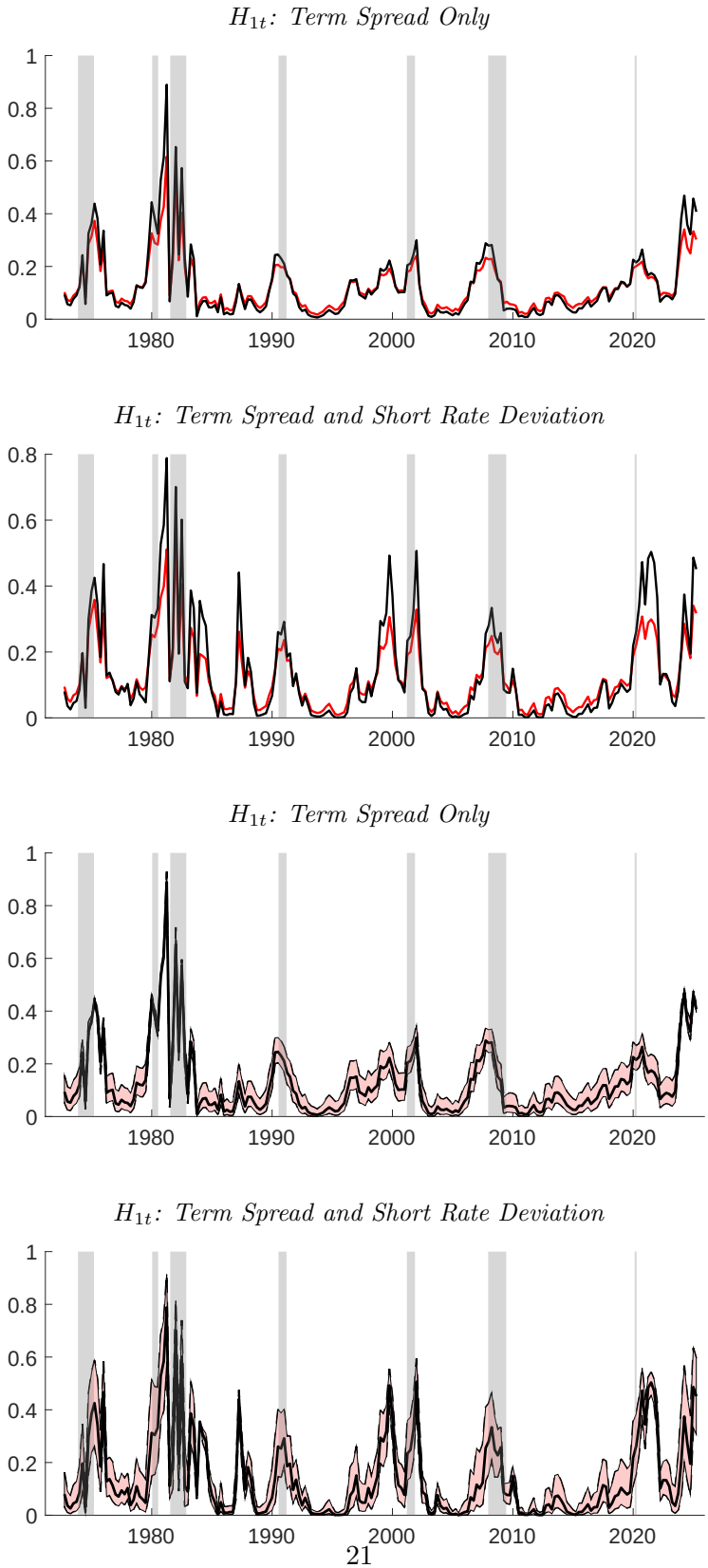
	NW						**LLSW-EWC**						**LLSW-NW**					
	g_{1t}	g_{2t}	g_{3t}	$g_{1(t-1)}$	$g_{2(t-1)}$	$g_{3(t-1)}$	g_{1t}	g_{2t}	g_{3t}	$g_{1(t-1)}$	$g_{2(t-1)}$	$g_{3(t-1)}$	g_{1t}	g_{2t}	g_{3t}	$g_{1(t-1)}$	$g_{2(t-1)}$	$g_{3(t-1)}$
2y	0.174	0.152	0.147	0.135	0.151	0.167	0.167	0.133	0.129	0.130	0.127	0.139	0.169	0.134	0.130	0.129	0.131	0.143
3y	0.168	0.155	0.145	0.136	0.151	0.165	0.160	0.136	0.127	0.133	0.128	0.138	0.162	0.135	0.125	0.131	0.133	0.146
4y	0.163	0.154	0.145	0.138	0.150	0.165	0.152	0.135	0.128	0.136	0.129	0.140	0.154	0.137	0.128	0.134	0.133	0.144
5y	0.156	0.159	0.143	0.141	0.151	0.161	0.146	0.139	0.127	0.137	0.128	0.138	0.145	0.139	0.128	0.139	0.132	0.142
6y	0.151	0.161	0.144	0.144	0.150	0.160	0.141	0.140	0.124	0.139	0.128	0.137	0.143	0.142	0.128	0.141	0.134	0.140
7y	0.149	0.159	0.141	0.146	0.150	0.157	0.137	0.141	0.124	0.142	0.127	0.134	0.138	0.144	0.126	0.143	0.134	0.138
8y	0.146	0.160	0.141	0.145	0.151	0.157	0.137	0.142	0.126	0.144	0.128	0.132	0.135	0.145	0.127	0.143	0.133	0.138
9y	0.146	0.160	0.145	0.146	0.151	0.158	0.136	0.140	0.129	0.144	0.130	0.133	0.133	0.143	0.130	0.143	0.132	0.137
10y	0.144	0.160	0.147	0.149	0.151	0.154	0.134	0.138	0.129	0.144	0.128	0.133	0.134	0.139	0.129	0.142	0.132	0.137

SA-3 Probability of Recession: Additional Results

In the main text, we report results for the conditional probability that $H_{2,t} = 1$. Figure SA.1, instead, shows the results for $H_{1,t} = 1$. The fitted probabilities are lower than for $H_{2,t}$ since $\{H_{1,t} = 1\} \Rightarrow \{H_{2,t} = 1\}$. The top two charts of Figure SA.1 compare the full-sample (leave-one-out) standard probit estimator (red line) with the bootstrap-bias corrected (leave-one-out) probit estimator (black line). Across both specifications, 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. The bottom two charts in Figure SA.1 show the associated 68% pointwise confidence intervals.

Figure SA.1. Probability of Recession in Four Quarters

This figure plots the fitted values of $P(H_{1,t} = 1|\, x_t)$ 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:Q3–2024:Q1.



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.

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

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

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.

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