University of Connecticut

Department of Economics Working Paper Series

Sieve Bootstrap Approach to Robust Term Premia Analysis

by

Jungbin Hwang
University of Connecticut

Feifan Wang
University of Connecticut

Working Paper 2025-10
September 2025

365 Fairfield Way, Unit 1063
Storrs, CT 06269-1063
Phone: (860) 486-3022
Fax: (860) 486-4463
http://www.econ.uconn.edu/

Sieve Bootstrap Approach to Robust Term Premia Analysis

Jungbin Hwang[1], Feifan Wang[2]

Department of Economics,
University of Connecticut, CT, USA

Abstract

Robust inference in bond predictive regressions faces challenges due to strong time-series persistence and unknown cross-sectional factor structures in the bond yield vector. These difficulties are particularly pronounced in analyzing the spanning hypothesis, which tests whether factors beyond the first three principal components (PCs)—level, slope, and curvature—improve bond return predictability. To address this, we develop a novel nonparametric sieve bootstrap approach for multivariate bond yield data with different maturities. Our method provides accurate size and improved power performance in bond predictive regression, compared to existing bootstrap inference procedures for the spanning hypothesis. Revisiting Cochrane and Piazzesi (2005)'s return-forecasting factor, we find strong evidence of its predictive power beyond the first three PCs for bond excess returns in most sample periods after the 1960s. However, we find that these predictive gains significantly decline when the sample period extends to include recent years after 2019.

Keywords: Sieve bootstrap, Term structure of interest rates, Predictive regression, Spanning hypothesis

JEL: C12, C15, E43, G12

1. Introduction

The term structure of interest rates plays a critical role in financial market analysis, offering key insights into economic conditions, guiding monetary policy decisions, and shaping

[1]Email: jungbin.hwang@uconn.edu
[2]Email: feifan.wang@uconn.edu

investment strategies. One of the key subjects in term structure analysis is bond risk premia, which refers to the difference between current long-term interest rates and the expected average of future short-term rates. Therefore, obtaining reliable predictions of future short-term rates is a crucial step in the empirical analysis of the term structure. This naturally leads to an important question: can the current yield curve alone provide enough information to forecast future short-term rates, or do we need to include other factors to better understand what affects risk premia?

This question is closely tied to the spanning hypothesis, which posits that under certain assumptions, the information contained in the current yield curve fully spans the relevant set of factors necessary for forecasting future interest rates and estimating risk premia. Empirical research finds that the first three principal components (PCs) of yields—commonly interpreted as level, slope, and curvature—serve as a concise summary of the yield curve (Litterman and Scheinkman, 1991). However, recent studies challenge the sufficiency of these three PCs, highlighting that predictive regressions of bond returns on various predictors, even when controlling for information in the current yield curve, often identify variables with additional predictive power (Cooper and Priestley, 2008; Ludvigson and Ng, 2009, 2010; Bansal and Shaliastovich, 2013; Joslin et al., 2014; Cieslak and Povala, 2015).

Despite these findings, several challenges arise in achieving robust estimation and inference in predictive regressions under the spanning hypothesis. One key issue is the unreliability of standard t-tests for testing the null hypothesis in bond predictive regressions, mainly due to the characteristics of bond yield data. Many studies highlight that predictors used in these regressions, such as bond principal components, are highly persistent. This persistence leads to finite-sample biases and size distortions, as discussed by Goetzmann and Jorion (1993), Stambaugh (1999), Valkanov (2003), Campbell and Yogo (2006), Wei and Wright (2013) and Bauer and Hamilton (2018). This issue becomes more challenging when explanatory variables show trends over the sample period. Additionally, predictive regressions often use overlapping data, such as annual excess bond returns analyzed with monthly observations. This overlap further undermines the reliability of standard errors and complicates the interpretation of regression R^2.

In addition to the time-series dependence, the cross-sectional dependence of yield curves across different maturities introduces further complexities, which is related to the identification and estimation of the unknown factor structure among different bond maturities. Recent studies by Crump and Gospodinov (2022, 2024) show that, even in the absence of strong serial correlation, certain features of maturity-ordered assets can cause standard principal component analysis to underestimate the true dimensionality of the underlying factor space.

To address these limitations, the literature has proposed bootstrap-based methods as alternatives to conventional inference techniques. Bauer and Hamilton (2018) introduced a parametric bootstrap approach to improve inference in predictive regressions under the spanning hypothesis. While their method addresses aforementioned challenges in bond predictive regressions, it relies heavily on the first three PCs of yields. This reliance introduces the risk of misspecification in dynamic settings where additional factors may influence risk premia and bond return dynamics. A recent work by Crump and Gospodinov (2024) proposed a model-free moving block bootstrap procedure that ensures robustness against unaccounted time-series and cross-sectional dependencies, as well as conditional heteroskedasticity. Despite this robustness, their approach is highly sensitive to the choice of block size, which can significantly affect its performance and inference outcomes. These limitations motivate the development of a new bootstrap methodology that combines the flexibility of model-free approaches with the theoretical rigor needed for robust inference in predictive regressions.

In this paper, we develop a novel nonparametric bootstrap method—the sieve bootstrap—to address the econometric challenges associated with the estimation and inference in predictive regressions. The core idea of the sieve bootstrap is to approximate an infinite-order vector autoregressive (VAR) model with a sequence of finite-order VAR models, where the lag order increases with the sample size but at a much slower rate than the sample size. This approach avoids reliance on a fixed finite-dimensional model by approximating an infinite-dimensional nonparametric model using a sequence of finite-dimensional parametric models. The sieve bootstrap method was first introduced by Bühlmann (1997), who formally established its asymptotic theory for univariate time series. In this paper, we extend the univariate sieve bootstrap approach of Bühlmann (1997) to multivariate time series to address both the time-series and cross-sectional dependence in the bond yield vector. Instead of directly resampling yields or prices, our method focuses on resampling excess returns along with a single, longest-horizon forward rate. This approach aligns with the "deconstruction"process of the yield curve proposed by Crump and Gospodinov (2024), which effectively mitigates strong time-series dependence often observed in yield data.

To implement our procedure, we use the Akaike Information Criterion (AIC) to select the lag order for the finite-order VAR model fitted to the transformed yield curve data. The finite-order VAR coefficients are estimated using the Yule-Walker method. After estimating the model, we compute the centered residuals and generate bootstrap samples of the whole yield curve recursively. This sieve bootstrap procedure allows us to evaluate the properties of predictive regressions under the spanning hypothesis while addressing both the time-series and cross-sectional dependence without assuming any restrictive parametric structure.

We evaluate the performance of our proposed bootstrap method by replicating key features of the Fama-Bliss data for one- through five-year zero-coupon bond prices. Specifically, we focus on the tent-shaped pattern observed in the predictive regression of annual excess returns on a linear combination of five forward rates, as identified by Cochrane and Piazzesi (2005). Our findings show that the sieve bootstrap more effectively captures both the dynamics of the yield curve and the cross-sectional dependence patterns essential for excess return forecasting, while avoiding sensitivity to the choice of the tuning parameter. Additionally, through extensive Monte Carlo simulation experiments, we show that our proposed bootstrap method achieves accurate size and high power performance in finite samples across various maturities in the bond predictive regression. Importantly, we find that the parametric bootstrap in Bauer and Hamilton (2018) exhibits limitations in finite samples, particularly with respect to conservative size and low power, further emphasizing the robustness and precision of our approach. Also, our sieve bootstrap is less sensitive than the moving block bootstrap in Crump and Gospodinov (2024), whose results often vary depending on the choice of block size.

In our empirical application, We revisit the existing empirical results for the spanning hypothesis in bond predictive regressions. Following Bauer and Hamilton (2018), we assess the significance of the return-forecasting factor proposed by Cochrane and Piazzesi (2005) using bootstrap procedures. Our analysis confirms that the Cochrane-Piazzesi (CP)'s return-forecasting factor exhibits strong predictive power for excess bond returns. For example, at 60-month maturity, the R^2 increases from approximately 25% in a regression on the first three PCs of yields to over 30% when the CP factor is included alongside the first three PCs during the early part of the sample, with starting years between 1965 and 1970. However, when the sample period is extended to include years after 2019, the predictive power of the CP factor diminishes, as reflected in the convergence of regression lines for rolling windows starting from 1983 onward. For instance, at 36-month maturity, the increase in R^2 driven by the inclusion of the CP factor narrows from approximately 5% in 1970 to less than 2% by 1983.

We further employ our sieve bootstrap method alongside the parametric bootstrap approach to evaluate the predictive power of the CP factor by revisiting the null hypothesis that the first three PCs fully encompass all relevant predictive information. Our findings show that the sieve bootstrap effectively captures the predictive dynamics of bond returns, particularly at longer maturities, and remains robust even during periods of structural change, outperforming the parametric approach in scenarios of evolving data characteristics.

This paper makes two key contributions. First, compared to parametric bootstrap methods, our sieve bootstrap does not require prior knowledge of the correct factor structure or the true pricing model that generated the data. This enhances its flexibility and applicability across a wide range of settings. Additionally, our method is robust to unaccounted forms of the time-series and cross-sectional dependence as well as the presence of conditional heteroskedasticity. Compared to nonparametric bootstrap methods such as the moving block bootstrap of Crump and Gospodinov (2024), our approach offers several practical advantages. Practically, the lag order in our model can be selected using information criteria such as AIC or BIC, providing a data-driven mechanism for model selection. Moreover, our sieve bootstrap can be implemented by the classical Yule-Walker method, which ensures the stationarity of the approximate VAR process in finite samples. This is an advantageous feature over moving block bootstrap methods in practice, which often require additional procedures to maintain stationarity in the bootstrapped data, e.g., Politis and Romano (1994).

Second, our empirical analysis provides a more accurate statistical inference for the spanning hypothesis by investigating the relevance of the Cochrane-Piazzesi (CP)'s return-forecasting factor—a linear combination of forward rates—which was not fully addressed in Crump and Gospodinov (2023, 2024). We find that the CP factor exhibits strong predictive power for excess bond returns and provides significant information beyond the first three PCs of yields. Furthermore, incorporating the CP factor consistently improves predictive accuracy, particularly for longer maturities. However, we also observe a notable decline in its effectiveness during the post-1983 period through 2024, a result not emphasized in prior studies. These findings underscore the importance of exploring additional predictors to enhance the robustness and accuracy of predictive regressions in bond markets.

The remainder of this paper is structured as follows. Section 2 explores the empirical motivation for revisiting the spanning hypothesis and its implications for bond market analysis. Section 3 discusses the econometric challenges of bond predictive regressions and the limitations of the parametric bootstrap approach. Section 4 introduces our sieve bootstrap methodology and provides details on how to implement it for multivariate bond yield data. Section 5 presents Monte Carlo experiments to highlight the advantages of our sieve bootstrap approach. Section 6 applies the sieve bootstrap to reexamine the CP factor, uncovering new insights into its gains in predicting bond excess returns beyond the level, slope, and curvature of the yield curve. Section 7 concludes and discusses directions for future research.

2. Empirical Motivation: Spanning Hypothesis

We first provide empirical motivation for our robust term premia analysis and formulate the corresponding statistical hypothesis. Let $P_{n,t}$ denote the log price of a pure discount bond with n maturities at date t. For our analysis, we use the Fama-Bliss data set, which is available from the CRSP database. This data set provides monthly observations of zero-coupon bond prices $P_{n,t}$ for maturities ranging from one to five years $n \in \{12, 24, 36, 48, 60\}$.

The associated log yield, denoted by $Y_{n,t}$, satisfies the simple relationship $P_{n,t} = -nY_{n,t}$. Also, we define the log forward rate, $f_{n,t}$, which corresponds to a one-period investment between dates $t+n-1$ and $t+n$ as:

$$f_{n,t} = P_{n-1,t} - P_{n,t}, \tag{1}$$

and the log forward rate $f_{n,h,t}$ for an h-period investment from $t+n-h$ to $t+n$:

$$f_{n,h,t} = P_{n-h,t} - P_{n,t}. \tag{2}$$

Given that $P_{0,t} = 0$ by construction, the one-period rate $Y_{1,t}$ is equal to $f_{1,t}$. Using the recursive nature of (1) and the definition of bond yields, we can also write that

$$P_{n,t} = -\sum_{i=1}^{n} f_{i,t}, \quad Y_{n,t} = \frac{1}{n}\sum_{i=1}^{n} f_{i,t}.$$

Following Cochrane and Piazzesi (2005, 2008), we define real risk-premia of n-maturity long bond as the excess h-period holding return:

$$rx_{n,h,t+h} = P_{n-h,t+h} - P_{n,t} + P_{h,t}. \tag{3}$$

Note that, with $h = 12$, $rx_{n,h,t+h}$ can be interpreted as borrowing at the one-year spot rate, buying a long bond with n-maturity, and sell it in one year. Thus, the use of $rx_{n,h,t+h}$ effectively removes the effects of inflation and the level of interest rates, allowing us to focus directly on real risk premia within the nominal term structure.

Our empirical question is grounded in the spanning hypothesis, a foundational theoretical argument in the term premia literature (Duffee, 2013). This hypothesis posits that, under certain assumptions about financial markets, the current yield curve fully spans the information set relevant for forecasting future interest rates and estimating risk premia (Bauer and Hamilton, 2018). To formalize the spanning hypothesis in an empirical context, our analysis employs the following bond predictive regression:

$$rx_{n,h,t+h} = \beta_{0n} + \beta_{1n}' F_{1t} + \beta_{2n}' F_{2t} + u_{t+h}, \tag{4}$$

for $n \in \{24, 36, 48, 60\}$, where $F_{1t} \in \mathbb{R}^{K_1}$ and $F_{2t} \in \mathbb{R}^{K_2}$.

The conventional spanning hypothesis assumes that the level, slope, and curvature of the yield curve encapsulate all the information available to investors at time t for predicting bond yields and excess returns across different maturities in future periods. Consequently, the first three PCs of the yield curve, captured in F_{1t} span the relevant information and are sufficient to model the dynamics of bond prices and yields. However, some studies suggest that the first three PCs of observed yields may not fully capture the relevant information. These studies argue that the first three PCs, F_{1t}, may not fully span all the relevant information and suggest that there may be an additional predictor, F_{2t}, that can improve forecasts of future yields and estimates of bond risk premia. This motivates the null hypothesis of interest: for each $n \in \{24, 36, 48, 60\}$,

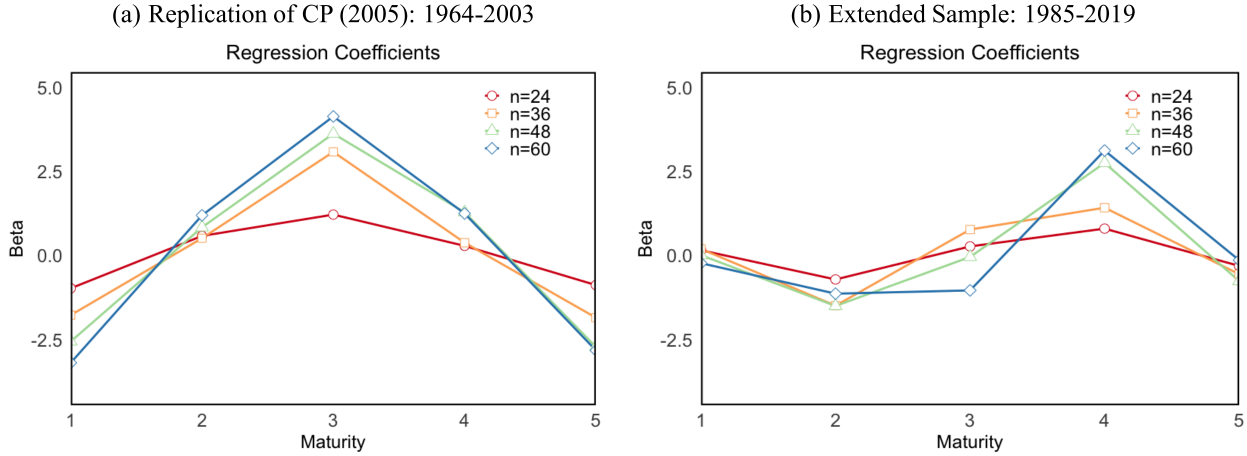
$$H_{0n} : \beta_{2n} = 0 \ . \tag{5}$$

What additional factor F_{2t} could be used for bond excess return prediction in (4)? A well-known finding related to testing the spanning hypothesis is provided by Cochrane and Piazzesi (2005, 2009, hereafter CP). The striking result in CP (2005) is that a single linear combination of the forward rates vector $\mathbf{f}_t := (Y_{12,t}, f_{24,h,t}, \ldots, f_{60,h,t})'$, which is unrelated to F_{1t}, can predict a substantial portion of $rx_{n,h,t+h}$ for all maturities. Specifically, CP (2005) run the following monthly predictive regressions with $h = 12$:

$$rx_{n,h,t+h} \quad = \gamma_n' \mathbf{f}_t + v_{n,h,t+h} = \gamma_{0,n} + \gamma_{1,n} Y_{12,t} + \gamma_{2,n} f_{24,h,t} + \cdots + \gamma_{5,n} f_{60,h,t} + v_{n,h,t+h}. \tag{6}$$

In Figure 1, we plot the estimated coefficients from (6) for two time periods. The first period, 1964–2003, is the same as CP (2005), and the corresponding results in the left panel of Figure 1 replicate their findings. The second period, 1985–2019, shown in the right panel of Figure 1, follows Bauer and Hamilton (2018) but extends their sample to include

Figure 1: Coefficient Estimates for Tent-shaped Regression



Notes: This figure displays the estimated coefficients $\hat{\gamma}_{i,n}$ across different maturities $n \in \{24, 36, 48, 60\}$ in tent-shaped regressions. The left chart of this figure represents the replication of the tent-shaped pattern reported by CP (2005) for the period 1964–2003, while the right chart of this figure shows the estimated coefficients for the period 1985–2019.

observations from 2016 to 2019. In both periods, we can check common patterns in the regression coefficients vector γ_n for all $n \in \{24, 36, 48, 60\}$, where a clear 'tent-shaped' pattern is evident in the first period, 1964–2003. CP (2005) argue that the presence of common patterns in (6) strongly supports the relevance of higher-order principal components, such as the fourth and fifth, and they can play a significant role in predicting bond returns.

Motivated by the common pattern in the estimated regression coefficients in (6), CP (2005) propose a single factor F_{2t} to summarize the expected excess returns across all maturities. Specifically, the single-factor regression can be written as:

$$rx_{n,h,t+h} = b_n F_{2t} + v_{n,h,t+h},$$

where $F_{2t} = a_0 + a_1 Y_{12,t} + \cdots + a_5 f_{60,h,t}$. By construction, the coefficients b_n and $a = (a_0, \ldots, a_5)'$ are not uniquely identified. To address this, CP (2005) impose the normalization condition that the average value of b_n equals one, i.e., $(1/4)\sum_{n=2}^{5} b_n = 1$. This indicates that the return forecast factor F_{2t} can be simply estimated by running the following regression:

$$\overline{rx}_{h,t+h} := \frac{1}{4}\sum_{n=2}^{5} rx_{n,h,t+h} = a_0 + a_1 Y_{12,t} + a_2 f_{24,h,t} + \cdots + a_5 f_{60,h,t} + \overline{v}_{h,t+h}, \tag{7}$$

from which we obtain $\hat{F}_{2t} = \hat{a}_0 + \hat{a}_1 Y_{12,t} + \cdots + \hat{a}_5 f_{60,h,t}$. Plugging the estimated CP factor $\hat{F}_{2t}$ into the original bond predictive regression in (4), we obtain the following bond predictive regression:

$$rx_{n,h,t+h} = \beta_{0n} + \beta_{1n}' F_{1t} + \beta_{2n}' \hat{F}_{2t} + u_{t+h} \tag{8}$$

for $n \in \{24, 36, 48, 60\}$. CP (2005) show that including the return-forecasting factor significantly increases the R^2 of the predictive regression in the first sample period (1964–2003). However, Bauer and Hamilton (2018) find that the R^2 gain from higher-order PCs diminishes notably in the second sample period (1985–2016), arguing that the predictive gain from CP's return-forecasting factor depends on the sample period. Readers are referred to Section 6, which provides detailed results for R^2 in regression (8).

A natural question now arises to complete the empirical analysis on the spanning hypothesis: are the estimated coefficients $\hat{\beta}_{2n}$ statistically significant? This question can be addressed by testing the null hypothesis of (5). Unfortunately, the standard t-test cannot be used to evaluate this simple null hypothesis within our bond predictive regression framework. The reasons for this limitation will be discussed in detail in the next section.

3. Econometric Issues on Bond-Predictive Regressions

3.1. Challenge in bond predictive regression inference

To illustrate the limitations of the standard t-test approach, we consider the case $K_1 = K_2 = 1$. Then, the standard t-statistic for testing β_{2n} in the (infeasible) bond predictive regression (4) can be expressed as:

$$t_T := \frac{\hat{\beta}_{2n} - \beta_{2n}}{\sqrt{se(\hat{\beta}_{2n})}} = \frac{\sum_{t=1}^{T} u_{t+h}\tilde{F}_{2t}}{\sqrt{s^2 \sum_{t=1}^{T} \tilde{F}_{2t}^2}}$$

where $\tilde{F}_{2t} = \ddot{F}_{2t} - \hat{A}_T \ddot{F}_{1t}$ with $\hat{A}_T = (\sum_{t=1}^T \ddot{F}_{2t}\ddot{F}_{1t})(\sum_{t=1}^T \ddot{F}_{1t}^2)^{-1}$, $\ddot{F}_{1t} = F_{1t} - \bar{F}_1$, and $\ddot{F}_{2t} = F_{2t} - \bar{F}_2$. s^2 is the standard error of regression which is computed as $(T - K_1 - K_2)^{-1} \sum_{t=1}^{T-h} (rx_{n,h,t+h} - \hat{\beta}_{0n} - \hat{\beta}_{1n}\ddot{F}_{1t} - \hat{\beta}_{2n}\ddot{F}_{2t})^2$, and $\hat{\beta}_{1n}$ and $\hat{\beta}_{2n}$ are ordinary least squares (OLS) estimates from (4). The validity of the standard Wald inference depends on whether t_T approximates well standard normal $N(0,1)$ distribution in finite-samples. This can be guaranteed if the time series data $rx_{n,h,t+h}$, F_{1t}, and F_{2t} are stationary and ergodic, and thus the estimate $\hat{A}_T$ converges to the population value $A = E[F_{2t}(F_{1t} - E[F_{1t}])](E[(F_{1t} - E[F_{1t}])^2])^{-1}$. Below, we discuss the key characteristics of bond predictive regression that drive the poor finite-sample performance of the standard t-test under normal approximations.

The first point closely follows those presented in Bauer and Hamilton (2018, Section 1), which point out that bond principal components and original bond yield data are highly persistent. The persistent predictors, F_{1t} and F_{2t}, result in poor finite-sample approximations for the standard first-order normal approximation of t_T. Motivated by this issue, Bauer and Hamilton (2018) derive a more accurate finite-sample approximation for t_T by using a local-to-unity specifications, e.g., Phillips (1988) and Cavanagh et al. (1995). Specifically, assume that $h = 1$ and u_{t+1} is a white noise and consider the case where $F_{1t} \in \mathbb{R}$ and $F_{2t} \in \mathbb{R}$ follow the local-to-unity AR(1) processes:

$$
\begin{aligned}
F_{1,t+1} &= \mu_1 + \left(1 + \frac{c_1}{T}\right) F_{1t} + \varepsilon_{1,t+1} \; ; \\
F_{2,t+1} &= \mu_2 + \left(1 + \frac{c_2}{T}\right) F_{2t} + \varepsilon_{2,t+1},
\end{aligned}
$$

where c_1 and c_2 are local-to-unity parameters. Assuming that $(\varepsilon_{1t}, \varepsilon_{2t})'$ is martingale difference sequence (m.d.s.) with $cov(\varepsilon_{1t}, \varepsilon_{2t}) = 0$, and ε_{2t} are independent of u_{t+1}, Bauer and Hamilton (2018, Appendix A.2) show that

$$
A_T = \frac{\sum_{t=1}^T \left(F_{1t} - \bar{F}_1\right)\left(F_{2t} - \bar{F}_2\right)}{\sum \left(F_{2t} - \bar{F}_2\right)^2} \Rightarrow \frac{\sigma_2 \int_0^1 J_{c_1}^\mu(\lambda) J_{c_2}^\mu(\lambda) d\lambda}{\sigma_1 \int_0^1 \left[J_{c_1}^\mu(\lambda)\right]^2 d\lambda} := \left(\frac{\sigma_2}{\sigma_1}\right) \tilde{A},
$$

where $Var(\varepsilon_{it}) = \sigma_i$, and $J_{c_i}^\mu(\cdot) = J_{c_i}(\cdot) - \int_0^1 J_{c_i}(r)dr$ are demeaned Ornstein-Uhlenbeck (OU) processes with $J_{c_i}(r) = c_i \int_0^r \exp(c_i(r-s))W_i(s)ds + W_i(r)$ for $i \in \{1, 2\}$, where $W_1(\cdot)$ are $W_2(\cdot)$ are independent standard Brownian motions. This result implies that even with a very large sample size T, the random variable $\hat{A}_T$, which is a component of the multiple regression coefficient $\hat{\beta}_{2n}$, does not converge to its true value A. Instead, it converges to a random variable $(\sigma_2/\sigma_1)\tilde{A}$ which is associated with the spurious regression coefficient between two independent I(1) processes F_{1t} and F_{2t}. Consequently, the finite-sample null distribution of the t-statistic can be very different from the standard normal random variable, as the local-to-unity asymptotics leads to

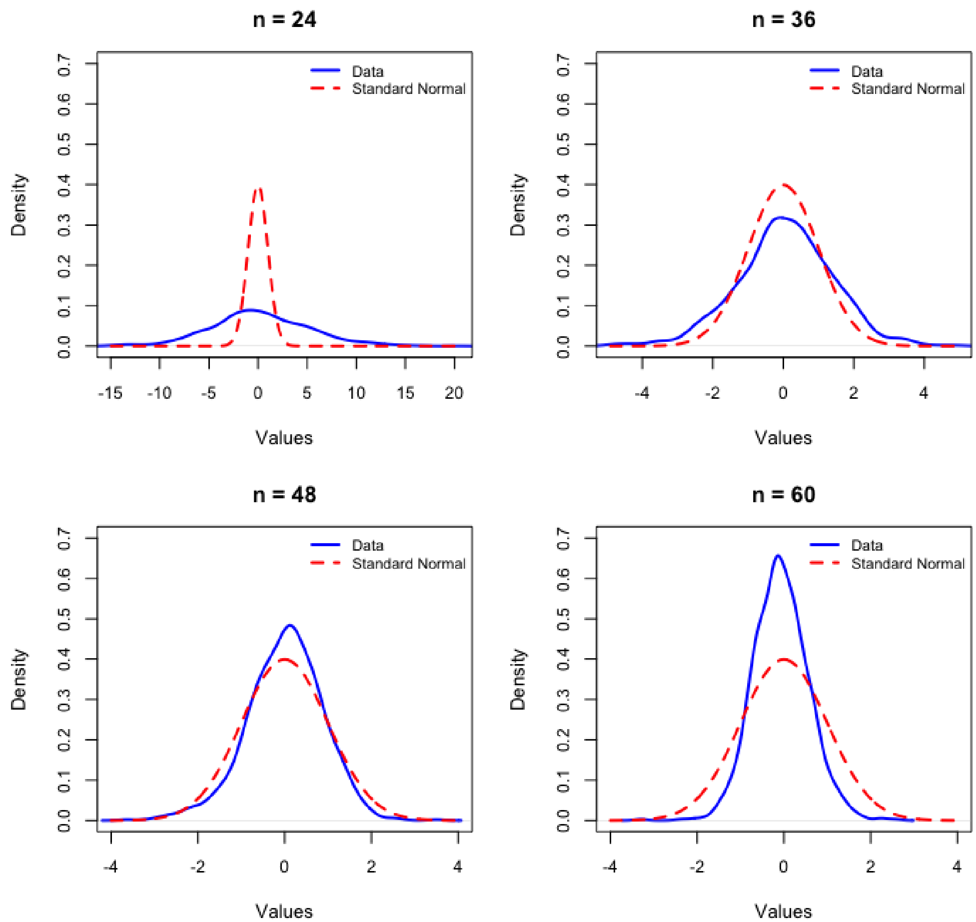
$$t_T \Rightarrow \delta Z_1(c_1, c_2, A) + \sqrt{1-\delta^2} Z_0,$$

where $\delta = corr(u_{t+1}, \varepsilon_{1t})$, $Z_0 \sim N(0,1)$, and $Z_1(c_1, c_2, A)$ is a functional of $J^{\mu}_{c_1}(\cdot)$, $J^{\mu}_{c_2}(\cdot)$, A, and $W_1(\cdot)$. As long as $\delta \neq 0$, i.e., the innovation of current yield curve's main PCs, F_{1t}, are correlated with the forecast error u_{t+1}, the limiting behavior of t_T can be largely deviated from $N(0,1)$. The result implies that the t-test for spanning hypothesis is prone to over-rejection in small samples, even when the null hypothesis $H_0 : \beta_2 = 0$ is true. Bauer and Hamilton (2018) show that the over-rejection of the null hypothesis can increase when the additional predictor F_{2t} has non-zero trending $\mu_2 \neq 0$. This non-zero trend can substantially amplify the aforementioned small-sample issues, leading to poorly sized tests and spurious rejections of the spanning hypothesis.

Another important issue in empirical bond predictive regressions arises from the use of overlapping data when studying annual excess bond returns with monthly observations. This issue applies to (4) and (6), which investigate one-year excess returns with $h = 12$, leading to $E\left(u_t u_{t-v}\right) \neq 0$ for $v \in \{0, \ldots, h-1\}$. The overlapping observations introduce an MA$(h-1)$ structure in the error terms of (4) and (6). In the presence of highly persistent F_{2t}, the product variable $u_{t+h}\tilde{F}_{2t}$ will derive a more complex limiting distribution for t_T, as shown by Müller (2014) and Xu (2020). The issue can leads to substantially misleading empirical conclusions such as false rejection of the null hypothesis, even after applying the well-known heteroskedasticity and autocorrelation consistent (HAC) correction to the denominator.

In addition to the time-series dependence in the yield data, the cross-sectional dependence between $rx_{n,h,t+h}$ for different maturities n can complicate the accurate identification of the true factor structure in the term structure of interest rates. Recent studies by Crump

Figure 2: The Failure of Standard t-Test



Notes: This figure illustrates the simulated standardized t-statistics from the predictive regression specified in (8) across four maturities $n \in \{24, 36, 48, 60\}$. The blue line depicts the empirical distribution of the simulated t-statistics under the null hypothesis $H_0 : \hat{\beta}_{2n,s} = \beta_{2n}$. The red dashed line represents the standard normal distribution for comparison. The simulations were conducted with $S = 1000$ replications.

and Gospodinov (2022, 2024) reveal that, even without strong serial correlation, certain characteristics of maturity-ordered assets may cause standard principle component analysis to underestimate the true dimensionality of the underlying factor space. Indeed, as we illustrate in the next subsection, imposing the true factor structure in the bond yield data may fail to reliably summarize the true cross-sectional dependence pattern in bond predictive regressions analysis.

To quantify these challenges in bond predictive regression inference, we conduct a Monte Carlo simulation using the data generating process (DGP) specified later in Subsection 5.1.

The goal of this simulation is to assess the finite-sample properties of the standard t-statistic under a setting where $\hat{\beta}_{2n}$ is close to its true value β_{2n}. We estimate the bond predictive regression model in (8) for maturities $n \in \{24, 36, 48, 60\}$ using a large-sample setting with $T = 10^6$. The OLS estimates of β_{2n}, which are nonzero, obtained from this large sample are treated as the population parameters for subsequent analysis. Next, we conduct $S = 1000$ Monte Carlo replications, each with a sample size of $T = 800$. For each replication s, we compute the OLS estimate $\hat{\beta}_{2n,s}$. The studentized t-statistic is then computed for the four maturities using the standard formula:

$$t_{2n,s} = \frac{\hat{\beta}_{2n,s} - \beta_{2n}}{\mathrm{se}\left(\hat{\beta}_{2n,s}\right)}.$$

Figure 2 presents the empirical distribution of the simulated t-statistic $t_{2n,s}$ for maturities $n \in \{24, 36, 48, 60\}$. The red dashed line represents the standard normal distribution, serving as a visual benchmark for comparison. For the 24-month maturity, the t-statistics exhibit extreme deviations, ranging from -15 to 20 and a heavy right tail extending to -20. This reflects significant positive skewness, consistent with the non-normal limiting distribution derived under local-to-unity asymptotics. The distortions diminish for longer maturities: the 36-, 48-, and 60-month panels show tighter ranges (-4 to 4), though residual excess kurtosis persists. Notably, the 60-month maturity still deviates visibly from the standard normal benchmark, particularly in the tails, underscoring the pervasive finite-sample distortions even for longer horizons.

The challenges discussed in this subsection—persistent predictors leading to non-standard asymptotics, overlapping observations inducing autocorrelated errors, and cross-sectional dependence complicating factor identification—undermine the reliability of standard t-tests in bond predictive regressions. Our simulation results confirm that conventional normal approximations fail to characterize the empirical distribution of t-statistics, resulting in systematic over-rejection of the null hypothesis. The severity of these distortions varies by maturity, with shorter maturities exhibiting more extreme deviations due to stronger persistence and local-to-unity behavior. Even for longer maturities, residual excess kurtosis and tail distortions persist, indicating that standard inference methods remain problematic across different bond horizons. These findings align with the theoretical results of Bauer and Hamilton (2018) and emphasize the need for more robust inference methods in bond return predictability studies.

3.2. Limitation of parametric bootstrapping

As an alternative econometric procedure to address the limitations of conventional t-statistic-based inference, the literature on bond predictive regression has proposed the bootstrap method (e.g., Bauer and Hamilton, 2018; Crump and Gospodinov, 2024). Since both the outcome variable ($rx_{n,h,t+h}$) and regressors (F_{1t}, F_{2t}) in the analysis are transformations of the original bond yield vector, $Y_t = (Y_{12,t}, \ldots, Y_{60,t})'$, the bootstrap methods aim to replicate these characteristics by carefully generating resampled yield vectors $(Y^*_{12,t}, \ldots, Y^*_{60,t})'$, which are then transformed to obtain bootstrapped versions of $(rx^*_{n,h,t+h})$ and (F^*_{1t}, F^*_{2t}). As a result, running bootstrapped versions of regressions provides a closer approximation of the true finite-sample properties of estimates from the original bond predictive regressions. The approach differs from the conventional bootstrap approach used in this literature, where resampled data are generated under the assumptions of the expectations hypothesis, e.g., Ludvigson and Ng (2009, 2010), and Greenwood and Vayanos (2014).

We describe the parametric bootstrap approach proposed by Bauer and Hamilton (2018). This method begins by computing the three principal components of the observed yields, denoted as $F_{1,t} = (PC_{1t}, PC_{2t}, PC_{3t},)'$ using the weighting vector $\hat{W}$, i.e., $F_{1,t} = \hat{W} Y_t$, where $\hat{W}$ is a 3×5 matrix with rows corresponding to the first three eigenvectors of the sample covariance matrix of Y_t. The eigenvectors are normalized such that $\hat{W}\hat{W}' = I_3$. Next, a VAR(1) model is estimated for $F_{1,t}$ using OLS:

$$F_{1,t} = \hat{\mu}_F + \hat{\Phi} F_{1,t-1} + \hat{\varepsilon}_{1,t} \tag{9}$$

for $t \in \{1, \ldots, T\}$. To generate bootstrapped samples of $F_{1,t}$, Bauer and Hamilton (2018) employ a residual bootstrap. The bootstrapped values $F^*_{1,t}$ recursively:

$$F^*_{1,t} = \hat{\mu}_F + \hat{\Phi} F^*_{1,t-1} + \varepsilon^*_{1,t},$$

where $\varepsilon^*_{1,t}$ are bootstrap residuals drawn independently from the set $\{\hat{\varepsilon}_{1,t}\}_{t=1}^T$ with replacement. The bootstrap process is initialized with $F_1^* = F_1$, the initial value of the observed sample. The bootstrapped yields are then derived using the relationship:

$$Y_t^* = \hat{W}' F_{1,t}^* + u_t^*, \tag{10}$$

for $t \in \{1, \ldots, T\}$, where $u_t^* \sim$ iid $N(0, \hat{\sigma}_u^2)$, and $\hat{\sigma}_u^2$ is the sample standard deviation of the errors $\hat{u}_t$. After obtaining the bootstrapped yield (price vector) Y_t^* (P_t^*) for $t \in \{1, \ldots, T\}$, we can obtain the bootstrapped versions of $rx^*_{n,h,t+h}$ as well as (F_{1t}^*, F_{2t}^*). These bootstrapped samples are then used to evaluate the properties of regression statistics in (4) under the spanning hypothesis in (5).

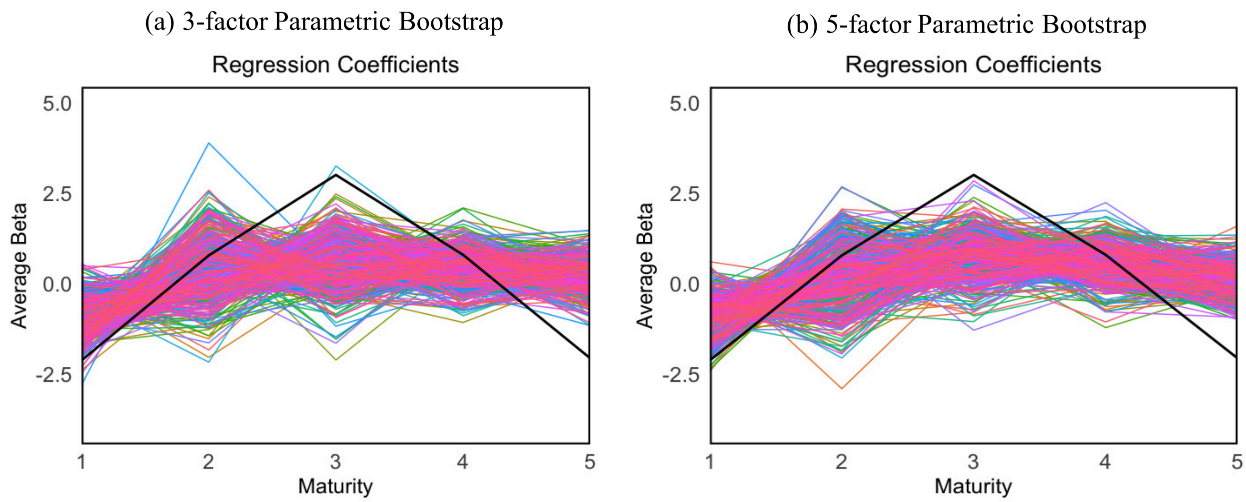
In Bauer and Hamilton (2018)'s setting, the generated data samples assume the spanning hypothesis holds, meaning no variables other than the PCs are used for prediction. This method addresses the first three econometric challenges outlined in the previous sections, including persistence, presence of trend and serial correlation problems. To evaluate the robustness of the parametric bootstrap method, we apply it to the bond data with maturities ranging from one to five years and run the bootstrap version of the regression in (6), i.e.,

$$rx^*_{n,h,t+h} \quad = \hat{\gamma}^*_{0,n} + \hat{\gamma}^*_{1,n} Y^*_{12,t} + \hat{\gamma}^*_{2,n} f^*_{24,h,t} + \cdots + \hat{\gamma}_{5,n} f^*_{60,h,t} + \hat{v}^*_{n,h,t+h}. \tag{11}$$

A robust bootstrap method is expected to replicate the characteristic tent-shaped pattern observed in Figure 1. However, our findings in the left panel of Figure 3, which plots 399 bootstrapped coefficients of the vector $(\hat{\gamma}^*_{1,n}, \ldots, \hat{\gamma}^*_{5,n})'$, indicate that the parametric bootstrap fails to capture this distinct pattern. Furthermore, the right panel of Figure 3 shows that even increasing the number of PCs from three to five does not resolve this issue.

Our replication of the parametric bootstrap approach raises concerns about the robustness of the procedure in Bauer and Hamilton's (2018) method. The method relies on a three-factor model as the true structure. However, using overlapping data induces cross-sectional dependence, complicating the identification of the true factor structure in the term structure of interest rates (Crump and Gospodinov, 2024). If the true factor structure differs

Figure 3: Resampling the Tent Shape using Bauer and Hamilton (2018)



Notes: This figure presents results using the parametric bootstrap method proposed by Bauer and Hamilton (2018), applied to the same dataset as CP (2005). The black line represents the tent-shaped pattern, while the multi-colored lines depict the bootstrapped coefficients of the vector $(\hat{\gamma}^*_{1,n}, \ldots, \hat{\gamma}^*_{5,n})'$. A total of $B = 399$ bootstrap replications were conducted.

from the one imposed in (10), their method may lack robustness and perform poorly in finite samples. Moreover, the parametric bootstrap approach assumes VAR(1) models to describe the dynamics of the factors. This assumption introduces additional concerns: even if the true model is a three-factor model, any deviation from VAR(1) factor dynamics in (9) would result in misspecification and further weaken the method's performance.

4. Sieve Bootstrapping for Yield Data

4.1. Sieve bootstrap approach

To precisely address the factor structure and issues related to the time-series and cross-sectional dependence in bond yield data, this section introduces a sieve bootstrap method and demonstrates its application to bootstrapping the bond predictive regression. The approach begins by transforming the original yield vector $(Y_{12,t}, Y_{24,t}, \ldots, Y_{60,t})'$ into less persistent stationary data. The process can be described as the "deconstruction" process of the yield curve in Crump and Gospodinov (2024). Specifically, we obtain $Z_t = (Z_{1,t}, \ldots, Z_{5,t})'$ for

$t \in \{2, \ldots, T\}$, where

$$
\begin{aligned}
Z_{1,t} &= P_{48,t} - P_{60,t} = f_{60,12,t}; \\
Z_{2,t} &= P_{36,t} - P_{48,t} - (P_{48,t-1} - P_{60,t-1}) = f_{48,12,t} - f_{60,12,t-1}; \\
Z_{3,t} &= P_{24,t} - P_{36,t} - (P_{36,t-1} - P_{48,t-1}) = f_{36,12,t} - f_{48,12,t-1}; \\
Z_{4,t} &= P_{12,t} - P_{24,t} - (P_{24,t-1} - P_{36,t-1}) = f_{24,12,t} - f_{36,12,t-1}; \\
Z_{5,t} &= -P_{12,t} - (P_{12,t-1} - P_{24,t-1}) = Y_{12,t} - f_{24,12,t-1}.
\end{aligned}
$$

By construction, each component of Z_t, except for the first, is formulated by taking the time-series difference between two neighboring forward rates. This ensures that the vector-valued stochastic process $\{Z_t\}_{t\in\mathbb{Z}}$, where $\mathbb{Z}$ is the set of all integers, is strictly stationary and preserves the original factor structure without relying on any parametric assumptions. The first component, $Z_{1,t} = f_{60,12,t}$, is less persistent than the original yield or price data, but it may still exhibit significant persistence depending on the data set and time period. In such cases, the prewhitening method, as in Andrews and Monahan (1992), can be applied to variables with strong serial correlation to improve the effectiveness of the bootstrap, e.g., Davison and Hinkley (1997) and Niebuhr et al. (2017). The detailed process of prewhitening $Z_{1,t}$ in our bootstrap procedure will be discussed later.

To proceed with the sieve bootstrap approach, we denote

$$E[Z_t] = (E[Z_{1,t}], E[Z_{2,t}], \ldots, E[Z_{N,t}])'$$

or equivalently $\mu = (\mu_1, \mu_2, \cdots, \mu_N)'$. The notation N represents the number of yield vectors, which equals 5 in our Fama-Bliss data. The autocovariance matrix functions are given by

$$\Gamma(k) = \mathrm{cov}(Z_t, Z_{t-k}) = E\left[(Z_t - \mu)(Z_{t-k} - \mu)'\right] \in \mathbb{R}^{N\times N}.$$

By construction, the symmetry $\Gamma(k) = \Gamma(-k)'$ holds for all $k \in \mathbb{Z}$. From the multivariate Wold decomposition theorem, the process $\{Z_t\}_{t\in\mathbb{Z}}$ can be written as an infinite-order causal vector moving average (MA) process, i.e.,

$$Z_t - \mu = \Theta(L)e_t = \sum_{\ell=0}^{\infty} \Theta_\ell \epsilon_{t-\ell}, \tag{12}$$

where $\Theta(z) = \sum_{\ell=0}^{\infty} \Theta_\ell z^\ell$ with $\Theta_0 = I_N$, and $\sum_{\ell=0}^{\infty} \|\Theta_\ell\|^2 < \infty$. Here, $\{e_t\}_{t\in\mathbb{Z}}$ is a white noise process with covariance matrix $E[e_t e_t'] = \Sigma$, where Σ is invertible. The MA(∞) representation allows us to express the autocovariance function as

$$\Gamma(k) = \sum_{\ell=0}^{\infty} \Theta_\ell \Sigma \Theta_{\ell+k}'.$$

We assume that the process $\{Z_t\}_{t\in\mathbb{Z}}$ is invertible so that it admits the following infinite-order vector autoregressive process:

$$A(L)(Z_t - \mu) = (Z_t - \mu) - \sum_{\ell=1}^{\infty} A_\ell (Z_{t-\ell} - \mu) = e_t, \tag{13}$$

where $A(z) = I_N - \sum_{\ell=1}^{\infty} A_\ell z^\ell$. By construction, the moving average and autoregressive lag polynomials are related through the equation $A(z)\Theta(z) = I_N$, which holds for all z. Moreover, each Θ_ℓ can be expressed in terms of the coefficients $\{\Theta_{\ell-i}\}_{i=1}^{\ell}$ and the autoregressive coefficient matrices $\{A_i\}_{i=1}^{\ell}$:

$$\Theta_\ell = \sum_{i=1}^{\ell} A_i \Theta_{\ell-i} \quad \text{for } \ell \geq 1.$$

Assumption 1. *$\{Z_t\}$ is a strictly stationary and purely nondeterministic stochastic process, where its autocovariance matrix function $\{\Gamma(k)\}_{k=-\infty}^{\infty}$ satisfies that $\sum_{k=-\infty}^{\infty}(1+|k|)^r \|\Gamma(k)\| < \infty$ for some $r \geq 0$. Additionally, there exists a constant $c > 0$ such that the spectral density matrix $f_Z(\cdot)$ of $\{Z_t\}_{t\in\mathbb{Z}}$ satisfies $\lambda_{\min}(f_Y(\lambda)) > c$ for all frequencies $\lambda \in (-\pi, \pi]$.*

Assumption 2. *$e_t = (e_{t,1}, e_{t,2}, \ldots, e_{t,N})'$ is an N-dimensional sequence of i.i.d. random vectors such that $E[e_t] = 0$, $E[e_t e_t'] = \Sigma \in \mathbb{R}^{N\times N}$ with $\Sigma > 0$, and $E[|e_{t,i} e_{t,j} e_{t,l} e_{t,s}|] < \infty$ for all $i, j, l, s \in \{1, \ldots, N\}$ and $t \in \mathbb{Z}$.*

Assumption 1 is analogous to Assumption (A) in Meyer and Kreiss (2015). The constant r in Assumption 1 determines the convergence rate of the VAR sieve approximation. The short memory condition in Assumption 1 guarantees that the spectral density $f_Z(\cdot)$ is continuously differentiable and uniformly bounded from above and below for all frequencies $\lambda \in (-\pi, \pi]$. Meyer and Kreiss (2015) show that any process satisfying Assumption 1 admits the one-sided representations of (12) and (13), where $\{e_t\}_{t\in\mathbb{Z}}$ is a strictly stationary, uncorrelated white noise process with a positive definite covariance matrix $E[e_t e_t'] = \Sigma$. Additionally, Assumption 1 guarantees that

$$\det(\Theta(z)) \neq 0 \quad \text{for all } |z| \leq 1 \text{ and } \sum_{\ell=1}^{\infty}(1+\ell)^r \|\Theta_\ell\| < \infty; \tag{14}$$

$$\det(A(z)) \neq 0 \quad \text{for all } |z| \leq 1 \text{ and } \sum_{\ell=1}^{\infty}(1+\ell)^r \|A_\ell\| < \infty \tag{15}$$

for $r \geq 0$, from Assumption 1. Therefore, the conditions in Assumption 1 ensure that $\Theta(z)$ and $A(z)$ are convergent for $|z| \leq 1$ and that the process $\{Z_t\}_{t\in\mathbb{Z}}$ is invertible, i.e., $\Theta^{-1}(z) = A(z)$.

The central idea of the sieve bootstrap method is to approximate an infinite-order VAR model with a sequence of finite p-order VAR models, where $p = p(T)$ increases with the sample size T, such that $p(T) = o(T)$. Instead of relying on a fixed finite-dimensional model, the sieve method approximates an infinite-dimensional nonparametric model using a sequence of finite-dimensional parametric models. This contrasts with the parametric bootstrap approach, such as that of Bauer and Hamilton (2018), which may be susceptible to misspecification in the dynamics of the factors. The sieve bootstrap approach was initially introduced by Bühlmann (1997), who established the consistency of the bootstrap for univariate time series. Here, we extend the single-dimensional sieve bootstrap to multivariate time series $Z_t \in \mathbb{R}^N$, as defined at the beginning of this section. By doing so, the sieve bootstrap approach can effectively handle both the time-series and cross-sectional dependence without assuming any restrictive parametric structure.

To extend this method to the multivariate context, we define $A(p) = (A_1, A_2, \ldots, A_p) \in \mathbb{R}^{N\times pN}$ as the matrix of the first p autoregressive parameter matrices in the VAR representation. The estimated p-order vector autoregressive coefficient matrix, denoted $\hat{A}(p) = [\hat{A}_{1,p}, \hat{A}_{2,p}, \dots, \hat{A}_{p,p}]$, can be obtained by fitting the observed data $\{Z_t\}_{t=2}^T$ using the Yule-Walker method (see Lütkepohl, 2005). Let $\hat{\Upsilon}(p)$ be a $(pN) \times (pN)$ matrix defined as

$$\hat{\Upsilon}(p) = [\hat{\Gamma}(j-i)]_{1 \le i,j \le p} = \begin{bmatrix} \hat{\Gamma}(0) & \cdots & \hat{\Gamma}(p-1) \\ \vdots & \ddots & \vdots \\ \hat{\Gamma}(-p+1) & \cdots & \hat{\Gamma}(0) \end{bmatrix},$$

where $\hat{\Gamma}(k)$ represents the k-th order sample autocovariance matrix:

$$\hat{\Gamma}(k) = \begin{cases} \frac{1}{T-k} \sum_{t=2}^{T-k} (Z_t - \bar{Z})(Z_{t+k} - \bar{Z})', & \text{for } k \ge 0, \\ \frac{1}{T+k} \sum_{t=2}^{T+k} (Z_t - \bar{Z})(Z_{t-k} - \bar{Z})', & \text{for } k < 0, \end{cases}$$

and $\bar{Z} = (T-1)^{-1} \sum_{t=2}^T Z_t$ is the sample mean of Z_t. The Yule-Walker estimate $\hat{A}(p)$ is then calculated as

$$\hat{A}(p) = [\hat{A}_{1,p}, \hat{A}_{2,p}, \dots, \hat{A}_{p,p}] = \left[\hat{\Gamma}(1), \dots, \hat{\Gamma}(p)\right] [\hat{\Upsilon}(p)]^{-1}.$$

It is well established in time series literature that the VAR process generated by $\hat{A}(p)$ using the Yule-Walker method is almost surely stationary. In addition, any finite-order VAR process generated by the Yule-Walker estimator is causal and invertible in finite samples. For further details and formal proof of the stationarity result, refer to Proposition 1 of Sun and Kaplan (2010), as well as Brockwell and Davis (1987, Chapter 8.1), Lütkepohl (2005, Chapter 3.3.4), and Reinsel (1993, Chapter 4.4).

Let $\hat{e}_{p,t}$ denote the residuals from the fitted VAR(p) model, defined as

$$\hat{e}_{p,t} = \left(Z_t - \bar{Z}\right) - \sum_{\ell=1}^{p} \hat{A}_{\ell,p} \left(Z_{t-\ell} - \bar{Z}\right),$$

for $t \in \{p+1, p+2, \dots, T\}$, with $\bar{Z} = (T-1)^{-1} \sum_{t=2}^T Z_t$. The covariance matrix Σ is estimated as

$$\hat{\Sigma}_p = \frac{1}{T-p}\sum_{t=p+1}^{T} \tilde{e}_{p,t}\tilde{e}'_{p,t},$$

where the centered residuals are given by $\tilde{e}_{p,t} = \hat{e}_{p,t} - \bar{\hat{e}}_p$, with $\bar{\hat{e}}_p = (T-p)^{-1}\sum_{t=p+1}^{T}\hat{e}_{p,t}$. The sieve bootstrap is then carried out as follows:

Step 0. (Prewhitening) Prewhiten the first component of Z_t, $f_{60,12,t}$, by fitting an AR(1) model and collecting the residuals $\tilde{f}_{60,12,t}$:

$$\tilde{f}_{60,12,t} = f_{60,12,t} - \left(\hat{\mu}_f + \hat{\rho}_f f_{60,12,t-1}\right), \quad \text{for } t = 2, \ldots, T,$$

where $\hat{\rho}_f$ is a bias-corrected OLS coefficient, as suggested by Kilian (1998). Replace $f_{60,12,t}$ with $\tilde{f}_{60,12,t}$ in Z_t, and resample $Z_t = (\tilde{f}_{60,12,t}, Z_{2,t}, \ldots, Z_{5,t})'$.

Step 1. Fit the p-order VAR model using the AIC criterion, and estimate $\hat{A}(p) = [\hat{A}_{1,p}, \hat{A}_{2,p}, \ldots, \hat{A}_{p,p}]$ following the Yule-Walker method. Compute the residuals of this model $\hat{e}_{p,t}$ for $t \in \{p+1, p+2, \ldots, T\}$ using the formula for $\hat{e}_{p,t}$ above.

Step 2. Center the residuals from Step 1, i.e., $\tilde{e}_{p,t} = \hat{e}_{p,t} - \bar{\hat{e}}_p$, and draw i.i.d. samples e_t^* from the following empirical distribution function:

$$\hat{F}_{e,T}(x) = \frac{1}{T-p}\sum_{t=p+1}^{T} \mathbf{1}[\tilde{e}_{p,t} \le x].$$

Step 3. Given initial values for $\{Z_t^*\}_{t=1}^{p}$, such as setting them equal to $\bar{Z}$, generate bootstrap replications of Z_t using the formula

$$Z_t^* - \bar{Z} = \sum_{\ell=1}^{p} \hat{A}_{\ell,p}\left(Z_{t-\ell}^* - \bar{Z}\right) + e_t^*. \tag{16}$$

for $t \in \{p+1, p+2, \ldots, T\}$. After generating a sufficiently large number of time periods, e.g., $10T$, discard the initial burn-in samples and use the last $(T-1)$ observations as the VAR sieve bootstrap samples $\{Z_t^*\}_{t=2}^{T}$. After the bootstrap procedure, the bootstrapped series $\tilde{f}_{60,12,t}^*$ is used to recursively reconstruct ("re-whiten") the bootstrapped forward rate $f^*_{60,12,t}$ as follows:

$$f^*_{60,12,t} = \hat{\mu}_f + \hat{\rho}_f f^*_{60,12,t-1} + \tilde{f}_{60,12,t}, \quad \text{for } t \in \{2, \ldots, T\},$$

where the initial value is set to $f^*_{60,12,1} = f_{60,12,1}$. The rewhitened $f^*_{60,12,t}$ then replaces $\tilde{f}_{60,12,t}$ in the first component of Z^*_t.

Step 4. Given the initial bond price vector $(P_{12,1}, P_{24,1}, \ldots, P_{60,1})'$, the bootstrapped prices can be recursively reconstructed using the bootstrapped sample $\{(Z^*_{1,t}, Z^*_{2,t}, \ldots, Z^*_{5,t})'\}_{t=2}^T$. Specifically, we generate the bootstrapped prices $\{(P^*_{12,t}, P^*_{24,t}, P^*_{36,t}, P^*_{48,t}, P^*_{60,t})\}_{t=2}^T$ as follows: Let $(P^*_{12,1}, P^*_{24,1}, \ldots, P^*_{60,1})' = (P_{12,1}, P_{24,1}, \ldots, P_{60,1})'$. For each $t \in \{2, \ldots, T\}$, we calculate:

$$\begin{aligned}
-P^*_{60,t} &= (Z^*_{1,t} + \cdots + Z^*_{5,t}) + (P^*_{12,t-1} - P^*_{60,t-1}) \\
P^*_{48,t} &= Z^*_{1,t} + P^*_{60,t}; \\
P^*_{36,t} &= Z^*_{2,t} + (P^*_{48,t-1} - P^*_{60,t-1}) + P^*_{48,t} \\
P^*_{24,t} &= Z^*_{3,t} + (P^*_{36,t-1} - P^*_{48,t-1}) + P^*_{36,t}; \\
P^*_{12,t} &= Z^*_{4,t} + (P^*_{24,t-1} - P^*_{36,t-1}) + P^*_{24,t}.
\end{aligned}$$

Step 5. From the reconstructed bootstrapped prices in the previous step, we compute forward rates $\{f^*_{n,12,t}\}_{t=1}^T$ for $n \in \{24, 36, 48, 60\}$, and 12-period holding excess returns $(rx^*_{n,12,t+12})$ for the entire bootstrapped sample using equations (2),(3), and compute bootstrapped data of testing factors F^*_{1t} and F^*_{2t}. These bootstrapped samples are then used to evaluate the properties of regression statistics in (4) under the spanning hypothesis in (5).

Our sieve bootstrap approach, summarized in Steps 1–5, resamples the transformed series Z_t, which effectively avoids strong time-series dependence commonly observed in yield data. By taking advantage of a flexible higher-order multivariate VAR(p) process in the bootstrapped process defined in (16), our approach is model-free and effectively mimics key properties of the data, including unknown factor structures, the time-series persistence, and cross-sectional dependence. This also allows our procedure to relax the strict assumptions of standard model-based bootstrap methods, such as Bauer and Hamilton's (2018) parametric bootstrap method.

As discussed earlier, a key feature of our bootstrap method is its reliance on only one persistent object—the longest-horizon forward rate, $\{f_{N,t}\}_{t=2}^{T}$ in the transformed yield data $Z_t = (Z_{1,t}, \ldots, Z_{N,t})'$. When the maturities in bond price vector has no gaps, i.e., $P_t = (P_{1,t}, P_{2,t}, \ldots, P_{N,T})$, Crump and Gospodinov (2024) note that the way how we transformed Z_t can be rooted in the following identity:

$$Z_{t+1} = (f_{N,t+1}, dr_{2,t+1}, \ldots, dr_{N,t+1})'$$

for $t \in \{1, \ldots, T-1\}$, where

$$dr_{n,t+1} \equiv \underbrace{P_{n-1,t+1} - P_{n,t}}_{:=r_{n,t+1}} - \underbrace{P_{n-2,t+1} - P_{n-1,t}}_{:=r_{n-1,t+1}} = f_{n,t} - f_{n-1,t+1}.$$

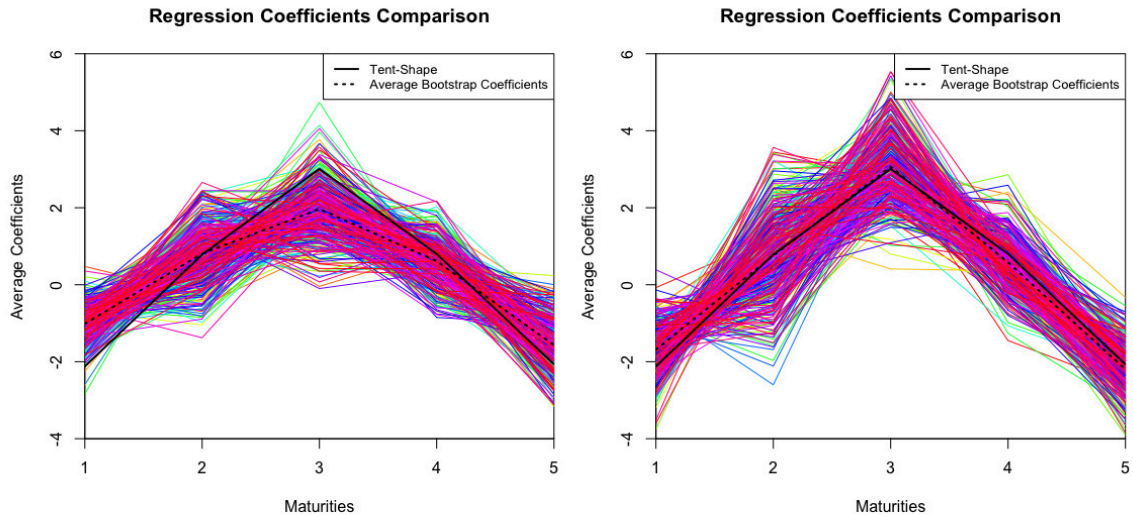
By construction, $dr_{n,t+1}$ is the difference return (or excess return), which can be obtained through a long-short trading strategy between periods t and $t+1$, an n-maturity bond is purchased, and $(n-1)$-maturity bond is shorted. Crump and Gospodinov (2024) propose resampling the economically motivated transformed data $(dr_{2,t+1}, \ldots dr_{N,t+1})'$, which possesses more appealing statistical properties, such as stationarity and moderate time-series dependence, along with a single, far-in-the-future forward rate $f_{N,t+1}$. The effectiveness of our transformed Z_t in the presence of gaps in maturities is confirmed in our empirical application in the next subsection.

4.2. Revisiting the tent-shape: a comparison of different bootstrap methods

To highlight the advantages of our proposed bootstrap method, we apply it to reexamine the tent-shaped predictive regression in (6) of Section 2. For comparison, we also implement the alternative bootstrap approach proposed by Crump and Gospodinov (2024), which employs the moving block bootstrap method for the transformed vector Z_t. For the choice of block size M, we follow their suggested rule-of-thumb formula $M = (TN)^{2/5}$, which gives

Figure 4: Revisiting the Tent Shape

(a) CG's Moving Block Bootstrap (b) Sieve Bootstrap



Notes: This figure presents results using the same dataset as described in subsection 3.2. The solid black line indicates the tent-shaped pattern, while the black dashed line represents the average value of the bootstrap coefficients. The multi-colored lines show the bootstrap estimates of the average regression coefficients for different maturities. The analysis is based on $B = 399$ bootstrap replications, with $T = 480$ and $N = 5$.

$M = 22$ for our sample period 1964–2003 with monthly observations $T = 480$ and $N = 5$. In our proposed method, we set the order of the VAR model to 10, which is consistent with popular data-driven criteria in VAR model such as AIC and BIC.

In addition to reporting the tent-shaped estimated regression coefficients for each bootstrap replication, we calculate the average value of the bootstrap coefficients, represented as a black dashed line in our graphs. The average line is compared with the tent-shaped line estimated from the original data, depicted as a solid black line, to evaluate the consistency and effectiveness of the methods.

In Figure 4, we observe that, compared to the parametric bootstrap approach presented in subsection 3.2, the nonparametric bootstrap methods, including our sieve VAR bootstrap and the moving block bootstrap of Crump and Gospodinov (2024), produce a similar contour to the original tent shape. Similarly, Figure A.5 in the Appendix illustrates the observed and sieve bootstrapped time series for yields, forward rates, and excess returns, showing that the sieve bootstrapped series closely align with the observed data. These findings indicate that our method more effectively captures the dynamics of the yield curve. Additionally, we modify the moving block size to 81, to replicate the result presented in Crump and Gospodinov (2023). With this adjustment, the performance of the moving block bootstrap improves compared to when the block size is set to the previous choice of $M = 22$. We find that our sieve VAR bootstrap approach induces similar results with tent-shaped patterns for flexible choices of p ranging from 7 to 12. These results further demonstrate that our method more effectively captures both the dynamics of the yield curve and the cross-sectional dependence patterns for excess return forecasting, without being sensitive to the choice of tuning parameter p.

In contrast to Crump and Gospodinov's (2024) moving block bootstrap, our method offers several practical advantages. Practically, the order of the VAR model, p, is easier and more straightforward to choose using popular methods such as AIC and BIC rules. Bühlmann (1997) shows that these data-dependent tuning parameters can have optimal properties depending on the true structure of the autoregressive process in the data. Bühlmann (1997) points out that selecting lag orders for an approximate VAR(p) process is more approachable for practitioners compared to choosing an optimal block length, which depends not only on the true dependence structure but also on the specific structure of the estimator to be bootstrapped.

Another advantageous feature is that the sieve VAR bootstrap approach employs the Yule-Walker method, which always ensures the stationarity of the approximate VAR(p) process in finite samples. This contrasts with moving block bootstrap-based methods, which require additional schemes to guarantee stationarity in the bootstrapped data, e.g., Politis and Romano (1994).

5. Monte Carlo Simulations

In this section, we present Monte Carlo simulation results to evaluate the performance of our proposed bootstrap method in bond predictive regression analysis. We compare it with existing approaches, such as the moving block bootstrap of Crump and Gospodinov (2024, hereafter CG) and the parametric bootstrap of Bauer and Hamilton (2018, hereafter BH). Our analysis begins by designing a data generation process for bond price data that closely aligns with the observed empirical patterns discussed in the Sections 2 and 3. All tables and figures in this section are included in the Appendix.

5.1. Data generation and calibration of population parameters

To begin our simulation analysis, we outline how to simulate the DGP of $P_t = (P_{12,t}, \ldots, P_{60,t})'$ to closely align with the empirical patterns of real bond data. The bond price data P_t is generated by first drawing the transformed variable Z_t, which follows a VAR(12) model:

$$Z_t - E[Z_t] = \sum_{\ell=1}^{12} A_{\ell,12} \left(Z_{t-\ell} - E[Z_t] \right) + \eta_t, \tag{17}$$

where each $A_{\ell,12}$ is 5×5 VAR coefficient matrix, $E[Z_t]$ is the population mean of Z_t, and η_t are i.i.d. Gaussian errors with variance-covariance matrix Σ_η. The parameters $\{A_{\ell,12}\}_{\ell=1}^{12}$, $E[Z_t]$, Σ_η from the Fama-Bliss data, which corresponds to CP (2005)'s sample period of 1964–2003 with monthly observations. After drawing simulated values Z_t from (17), we convert them back into simulated bond price vector P_t and the recursive relation shown in Step 4 of subsection 4.1. Note that the first element in Z_t, denoted as $f^R_{60,12,t}$, is the residual of the forward rate $f_{60,12,t}$,which satisfies that

$$f_{60,12,t} = \mu_f + \rho_f f_{60,12,t-1} + f^R_{60,12,t}, \tag{18}$$

where the parameters μ_f and ρ are calibrated using the same Fama-Bliss data.

To obtain the population parameters of the coefficient vector $\gamma_n = (\gamma_{1,n}, \gamma_{2,n}, \ldots, \gamma_{5,n})'$ for the tent-shaped regression specified in (6), we simulate the time series vector P_t as described above and use them to compute the dependent variable $rx_{n,h,t+12}$ along with the regressors $(Y_{12,t}, f_{24,12,t}, \ldots, , f_{60,12,t})$ for each $n \in \{24, 36, 48, 60\}$. Using a sufficiently large number of T, e.g., $T = 10^6$ in Monte Carlo samples, we run predictive regressions and treat the OLS estimates as the population parameters derived from the calibrated data generation process described in (6) and (18). Table A.1 presents the population parameters for the DGP used in our simulation analysis. To save space, we report only the first lag's coefficient matrix $A_{1,12}$ from the VAR(12) model. The population parameter vector γ_n in Table A.1 indicates that our DGP closely replicates the tent-shaped feature observed in the original sample.

The tent-shaped pattern of the population parameter γ_n is an advantageous feature of our Monte Carlo design, differentiating it from the approaches in BH (2018) and CG (2024). Their methods of simulating bond yield data rely on a simpler VAR(1) process that assumes a true three-factor structure. In contrast, our DGP does not assume a specific factor structure but instead incorporates more realistic bond yield dynamics within a VAR(12) framework. Consequently, our design reproduces the tent-shaped pattern in bond yield regressions, closely aligning with the empirical data.

5.2. Simulation evidence

Based on simulated samples for the tent-shaped regression (6) with $T = 800$, this section investigates the size and power properties of bootstrap inferences. The number of Monte Carlo replications is $S = 1,000$. For each s-th simulated samples, we compute the OLS estimates for $\gamma_{i,n}$, denoted by $\hat{\gamma}_{i,n,s}$, and the corresponding studentized t-statistic:

$$t_{i,n,s} = \frac{\hat{\gamma}_{i,n,s} - \gamma_{i,n}}{\mathrm{se}\left(\hat{\gamma}_{i,n,s}\right)}$$

for $i \in \{1, 2, \ldots, 5\}$, where $\mathrm{se}\left(\hat{\gamma}_{i,n,s}\right)$ is the Newey-West HAC standard error with a truncation lag parameter 12. Conditional on each simulated sample, we generate $B = 399$ bootstrap draws, where the corresponding studentized bootstrap t-statistic for the b-th replication is computed as:

$$t^*_{i,n,s(b)} = \frac{\hat{\gamma}^*_{i,n,s(b)} - \hat{\gamma}_{i,n,s}}{\mathrm{se}(\hat{\gamma}^*_{i,n,s(b)})}.$$

The p-value for the symmetric percentile-t bootstrap method can be then calculated as:

$$p^*_{|t_{i,n,s}|} = \frac{1}{B}\sum_{b=1}^{B} 1\{|t^*_{i,n,s(b)}| > |t_{i,n,s}|\},$$

where $1\{\cdot\}$ denotes the indicator function.

Empirical size and power are computed as follows. Note that the test statistic $t_{i,n,s}$ is centered around the non-zero population value of $\gamma_{i,n}$ presented in Table A.2. Thus, the finite-sample size is defined as the proportion of simulations in which the bootstrap p-value, calculated using the non-zero population value $\gamma_{i,n}$, exceeds the pre-specified nominal level α. The power, on the other hand, is defined as the rejection rate of the false null hypothesis $H_0 : \gamma_{i,n} = 0$, where the bootstrap p-value, with $\gamma_{i,n} = 0$, exceeds the nominal level α. To save space, we present results only for the $\alpha = 10\%$ nominal significance level, and results for $\alpha = 5\%$ are provided in the Appendix.

Table A.2 shows the size performance of three bootstrap methods: the parametric bootstrap approach introduced by BH (2018), the moving block bootstrap method developed by CG (2024), and the sieve bootstrap proposed in this paper. The results show that the sieve bootstrap yields empirical sizes close to the nominal 10% level across various maturities $n \in \{24, 36, 48, 60\}$ and regressors $(Y_{12,t}, f_{24,12,t}, \ldots, , f_{60,12,t})$ in the tent-shaped regression. For instance, the sieve bootstrap achieves empirical sizes ranging from 0.077 to 0.092 when testing the parameter $\gamma_{1,n}$ corresponding to $Y_{12,t}$, which are close to the nominal level. In contrast, the moving block bootstrap exhibits under-rejections, with empirical sizes often far below the nominal level; for example, the empirical size for testing $\gamma_{1,n}$ at the 3-year maturity ($n = 36$) is 0.003. The parametric bootstrap is overly conservative in finite samples, exhibiting more under-rejections than nonparametric bootstrap methods, particularly for parameters corresponding to regressors with long term interest rates, such as $f_{36,12,t}$ and $f_{60,12,t}$, where empirical sizes are almost zeros.

In Table A.3, we report the finite-sample power performance of the three bootstrap methods. Consistent with the size results, the sieve bootstrap demonstrates superior performance, achieving high power across most maturities and test variables. For instance, the sieve bootstrap attains power close to or equal to 1 in several cases, including testing parameters corresponding to $Y_{12,t}$ and $f_{48,12,t}$ at 2-year maturity. This indicates that our bootstrap approach exhibits favorable finite-sample properties for detecting true signals in bond predictive regression. The moving block bootstrap performs successfully in some cases (e.g., $f_{24,12,t}$ at 2 years) but shows substantially low power for other parameters (e.g., $f_{60,12,t}$ at 2 years).

The parametric bootstrap exhibits the weakest power among the three bootstrap methods, with values of finite-sample power significantly lower than those of moving block bootstrap approach in CG (2024) and the sieve bootstrap. We also simulate the empirical size and power results using the unstudentized t-statistics, and the results are reported in Tables A.4 and A.5. These results indicate similar findings, further reinforcing the finite-sample evidence of the favorable size and power properties of the sieve bootstrap method.

Overall, the nonparametric sieve bootstrap proposed in this paper demonstrates successful finite-sample performance, considering both size and power. The alternative moving block bootstrap and parametric bootstrap exhibit limitations in finite samples. For the parametric bootstrap, the size is severely under-rejected because its validity crucially depends on the assumption of the true structure of the factor space and the time-series dependence. Specifically, it assumes a three-factor structure and VAR(1) time series dynamics, so any deviations from this assumption are expected to result in poor performance. Our numerical results from Monte Carlo simulations confirm that these misspecifications in the parametric bootstrap lead to a substantial loss of finite-sample power.

Our findings also indicate that the moving block bootstrap proposed by CG (2024) can improve the performance of the parametric bootstrap, but its performance is highly sensitive to the choice of block size. In Tables A.10 and A.11 of the Appendix, we present results for different block sizes and observe that changes in block size can lead to significant variations in both size and power. The sensitivity to block size remains as the sample size increases. We also find that in some cases, the computation of the optimal block size suggested by the literature can exceed the available sample size T, further complicating its implementation in our bond yield data. In contrast, Table A.12 in the Appendix shows that our sieve bootstrap approach delivers consistent size performance across various lag orders p. This result indicates that the sieve bootstrap offers a simpler and more accessible approach to selecting the order of the VAR model, such as the AIC formula, consistently yielding more robust and reliable results.

6. Revisiting Spanning Hypothesis

In their application of the parametric bootstrap approach, BH (2018) present empirical analysis of the spanning hypothesis by revisiting the work of CP (2005). They argue that CP's return-forecasting factor lacks significant predictive power outside the specific periods chosen by CP (2005) and conclude that the CP factor is primarily driven by the first three PCs of yields, which capture all the relevant information for predicting bond returns. While their findings suggest that the level, slope, and curvature of the yield curve fully capture the predictive content of term structures, with no additional bond excess return prediction from the CP factor, their parametric bootstrap approach has limitations in bond predictive regressions. Specifically, the parametric bootstrap approach critically relies on the assumption of three PCs and VAR(1) dynamics for bond yields, which may not fully incorporate CP's return-forecasting factor in its bootstrap inference, raising concerns about the generalizability of the conclusions.

In this section, we first apply our sieve bootstrap method to reevaluate CP (2005)'s work on the empirical test of the spanning hypothesis. All tables and figures in this section are also included in the Appendix. Using monthly Fama-Bliss data from 1964 to 2024, with a 39-year rolling window starting in 1964, we first conduct a rolling R^2 analysis to evaluate the significance of CP (2005)'s return-forecasting factor. Figure A.6 presents the rolling R^2 results across four maturities ($n \in \{24, 36, 48, 60\}$) for regression (8), comparing models with and without the inclusion of the CP factor $\hat{F}_{2t}$. Our analysis shows a consistent increase in R^2 with the inclusion of the CP factor. For example, at $n = 60$, the value of R^2 rises from around 25% ((blue line, 3PCs only) to over 30% (red line, 3PCs + CP factor) during the early part of the sample, with starting years between 1965 and 1970. Also, Figure A.6 indicates that R^2 values increase with longer maturities, reaching their highest levels at $n = 48$ and $n = 60$, where R^2 consistently exceeds 30% during the late 1960s and early 1970s. Importantly, the predictive gains of the CP factor diminish over time, as evidenced by the convergence of the red and blue lines for rolling windows starting after 1983. For instance, at $n = 36$, the increase in R^2 driven by the inclusion of the CP factor narrows from approximately 5% in 1970 to less than 2% by 1983. Lastly, the green line in Figure A.6, representing regressions with only CP factor and a constant, highlights its standalone predictive power. During most rolling window periods with starting years between 1964 and 1980, we find that the R^2 values for the single CP factor dominate those for the three PCs, confirming CP (2005)'s original findings. However, the R^2 of the single CP factor regression begins to decline for starting years after 1983, which aligns with our earlier findings on the decreasing relevance of the CP factor in forecasting excess bond returns in more recent periods.

Following the rolling R^2 analysis, we employ the sieve bootstrap method, along with asymptotic t-inference and the parametric bootstrap method, to assess the predictive power of the CP factor. This allows us to revisit the null hypothesis in (5) that the first three PCs capture all the relevant predictive information. In Panel A of Table A.13, we report the results of HAC inference for the regression (8) using Newey-West standard errors with the truncation lag 12, over the same sample period as CP (2005): 1964–2003, The results imply that CP (2005)'s factor is highly statistically significant across all maturities, strongly rejecting the spanning hypothesis.

To obtain robust inference regarding the relevance of the CP factor, we also present results using bootstrap inferences in Panel A of Table A.13. For each bootstrap method, we perform $B = 399$ replications and report both studentized and unstudentized bootstrapped p-values under (5). The results indicate that the sieve bootstrap approach consistently produces statistically significant results for p-values constructed under both studentized and unstudentized statistics. This confirms the relevance of the CP factor during the sample period of CP (2005). In contrast, while the parametric bootstrap yields statistically significant p-values for $\hat{F}_{2t}$ with studentized statistic, we find that those with the unstudentized statistic are not statistically significant, indicating potential limitations of this method in providing robust inference.

In Panel B of Table A.13, we report the results for the later sample period, 1985–2019, revisiting the findings in BH (2018) with their parametric bootstrap approach. During this time, the predictive power of the regressions across different n, as indicated by R^2 values, generally decreases compared to the earlier sample in Panel A. However, the inclusion of the CP factor still improves predictive power, as reflected in the higher R^2 values across all maturities when the CP factor is included. In this later sample, the results of the parametric bootstrap show that neither the studentized nor unstudentized p-values are statistically significant across all maturities. In contrast, the sieve bootstrap method continues to deliver statistically significant results for both studentized and unstudentized p-values across all maturities, aligning with the R^2 gains from the CP factor.

To check whether our findings remain stable over different sample periods, we also compute bootstrap p-values, along with the differences in adjusted R^2 values, using the same rolling sample periods as those in Figure A.6. The results in Figure A.7 show that the p-values using the studentized statistic exhibit distinct patterns between the parametric and sieve bootstrap methods, with the red dashed line representing the 5% nominal significance level. Specifically, the parametric bootstrap shows greater variation over time and is statistically insignificant (at 5% nominal size) p-values more often. On the other hand, the p-values obtained from the sieve bootstrap approach imply consistently highly significant CP factor until the starting period around 1983, after which the CP factor substantially lose statistical significance as the rolling data period approaches more recent years. This result aligns with our previous R^2 analysis and suggests that better captures the dynamics of return-forecasting factor beyond the first three PCs. The unstudentized p-values in Figure A.8 provide similar insights. The parametric bootstrap p-values remain statistically insignificant throughout the entire rolling sample period, thus conflicting with the studentized statistic. In contrast, the sieve bootstrap follows a pattern similar to that of the studentized p-values.

In Figure A.9, we present the (adjusted) R^2 differences (green line) for regressions of excess bond returns ($rx_{n,12,t+12}$) on the three PCs of the yield curve, with and without the inclusion of the CP factor, across maturities ($n \in \{24, 36, 48, 60\}$). The black line represents the bootstrap mean values for the R^2 difference. By construction, the R^2 difference close to 0 indicate that the CP factor has limited predictive power in the bond predictive regression. For the parametric bootstrap (top panel of Figure A.9), the bootstrap means remain close to 0, failing to follow the sample pattern, which suggests potential limitations in its ability to capture the underlying structure. In contrast, the sieve bootstrap (bottom panel of Figure A.9) aligns more closely with the sample patterns, providing more reliable measures of sampling uncertainties in the R^2 difference estimates over time.

Given the strong predictive power of the CP factor for excess bond returns observed in most sample periods, we further investigate the reliability of the sieve bootstrap in capturing its predictive content by conducting a Monte Carlo simulation based on the bond predictive regression framework given in (8), which incorporates three PCs and the CP factor. The simulation setup follows the description of the failure of standard t-test in Subsection 3.1. After generating simulated samples, we select one simulated sample, with its corresponding OLS estimate denoted as $\hat{\beta}_{2n,sel}$, where $T = 800$, for further analysis. Both the BH and sieve bootstrap methods are applied to this selected sample, using $B = 1000$ replications. The sieve bootstrap method employs a lag order of $p = 20$.

The studentized t-statistics for the sieve and BH bootstrap methods are defined as follows:

$$t^{*\text{sieve}}_{2n,s(b)} = \frac{\hat{\beta}^{*\text{sieve}}_{2n,s(b)} - \hat{\beta}_{2n,sel}}{\text{se}\left(\hat{\beta}^{*\text{sieve}}_{2n,s(b)}\right)},$$

and

$$t^{*\text{BH}}_{2n,s(b)} = \frac{\hat{\beta}^{*\text{BH}}_{2n,s(b)} - \hat{\beta}_{2n,sel}}{\text{se}\left(\hat{\beta}^{*\text{BH}}_{2n,s(b)}\right)}.$$

In this context, the null hypothesis states that the bootstrapped estimates $\hat{\beta}^{*}_{2n,s(b)}$ are close to the simulated estimates $\hat{\beta}_{2n,sel}$. For comparison, we also report the results of the unstudenized tests for both bootstrap methods. Figure A.10 illustrates the distributions of the bootstrapped estimates $\hat{\beta}^{*}_{2n,s(b)}$ across maturities $n \in \{24, 36, 48, 60\}$ under studentized tests. The top panel shows the distributions obtained using the BH bootstrap method, while the bottom panel presents those derived from the sieve bootstrap method. The solid blue lines represent the bootstrap distributions, while the dashed red lines correspond to the simulated distributions for $\hat{\beta}_{2n,s}$, as described in Subsection 3.1.

The BH bootstrap distributions exhibit greater dispersion, particularly in the tails, indicating increased variability in the estimates. For instance, at $n = 24$, the BH bootstrap distribution demonstrates a pronounced right tail, suggesting a higher probability of extreme estimates. Similarly, at $n = 36$, the distribution appears less concentrated around the mean, with substantial spread in both tails. This increased dispersion suggests that the BH bootstrap is more sensitive to deviations from its underlying assumptions, such as the reliance on three PCs and a VAR(1) model for bond yields. Conversely, the sieve bootstrap distributions are more tightly centered around the simulated sample distributions, indicating greater stability and robustness. At $n = 24$, the sieve bootstrap distribution is symmetric and closely follows the simulated sample, exhibiting minimal tail dispersion. Likewise, at $n = 36$, the distribution remains well-clustered around the mean, showing little deviation from the simulated sample.

Figure A.11 presents the distributions of the bootstrapped estimates $\hat{\beta}^{*}_{2n,s(b)}$ under unstudentized tests. Similar patterns emerge, with the BH bootstrap distributions displaying greater variability and divergence from the simulated sample. For example, at $n = 24$, the BH bootstrap distribution exhibits noticeable skewness, while at $n = 36$, it remains less centered and shows pronounced tail behavior. In contrast, the sieve bootstrap distributions exhibit lower variability and improved consistency, closely following the simulated sample distributions across all maturities. These results underscore the sieve bootstrap's superior ability to approximate the true DGP.

In conclusion, our analysis confirms that, in most sample periods between 1964 and 2024, the CP's return-forecasting factor has strong predictive power for excess bond returns and thus provides significant information beyond the first three PCs of yields. This is supported by our R^2 analysis for rolling sample periods, which shows that incorporating the CP factor consistently improves predictive accuracy, particularly at longer maturities. However, its gains for bond return prediction substantially decline in sample periods starting in the years after 1983. The nonparametric bootstrap approach proposed in this paper further validates the statistical significance of the CP factor in most sample periods, including BH (2018)'s sample period (1985–2016) and CP (2005)'s (1964–2003). Our findings emphasize the importance of factors beyond the first three PCs in the current yield curve and the advantage of applying our robust sieve bootstrap methods to capture their dynamic predictive power for bond excess returns over time.

7. Conclusion and Future Research

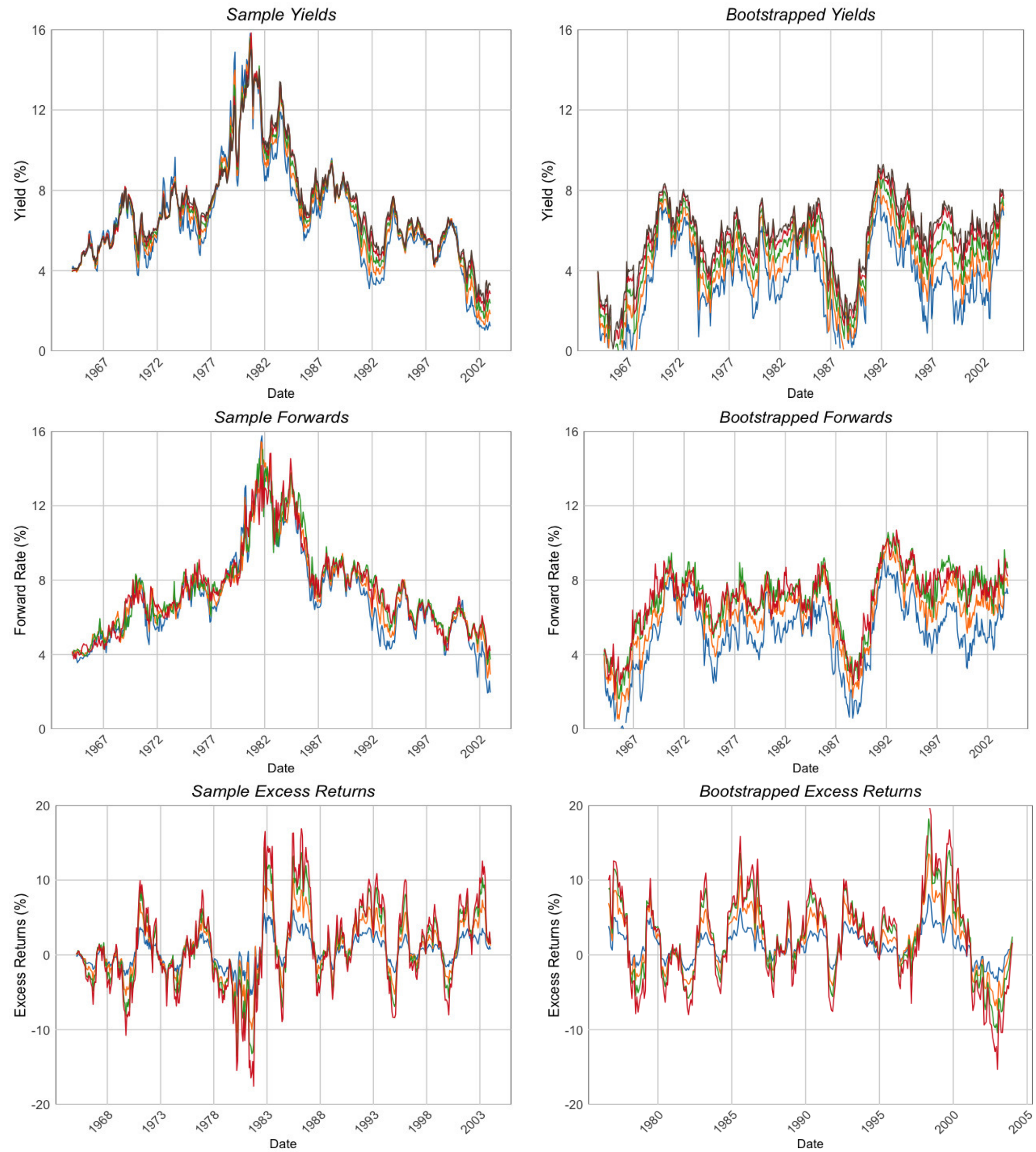
In this paper, we develop a sieve bootstrap approach as a robust and flexible approach to addressing challenges in the analysis of bond risk premia, particularly under the spanning hypothesis framework. By employing a nonparametric approach, we show that the sieve bootstrap effectively captures key data properties, such as cross-sectional dependencies and time-series dynamics, without relying on restrictive parametric assumptions. Our method overcomes limitations associated with existing bootstrap approaches, such as those of BH (2018) and CG (2024), by leveraging the Yule-Walker method to ensure stationarity and using data-driven criteria for model selection.

Our empirical findings confirm that the sieve bootstrap consistently delivers robust and statistically significant results, even in periods of structural change. Through Monte Carlo experiments, we validated its superiority in maintaining accurate empirical sizes and high power performance compared to alternative methods. Additionally, we reexamine the CP (2005)'s return-forecasting factor and found it to have strong predictive power across maturities, though its relevance diminishes in more recent periods.

While our analysis focuses on bond yield data with one- to five-year maturities at a one-year horizon, further extensions of the sieve bootstrap approach could include all maturities and varying investment horizons. Accommodating high-dimensional bond yield data, along with other relevant predictors such as macroeconomic or monetary fundamentals, could complicate bootstrap inference in finite samples. We leave this as a subject for future study.

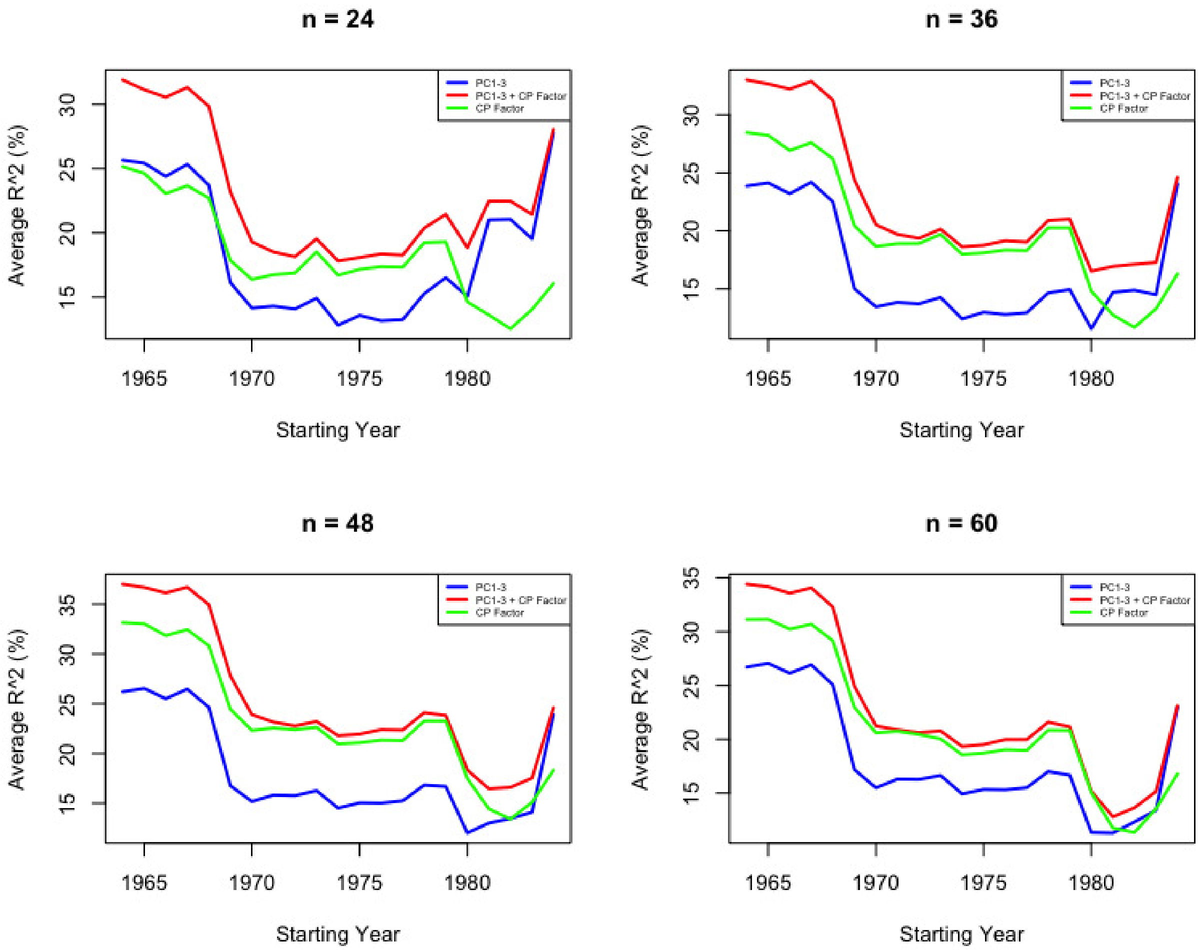
Appendix A. Figures and Tables

Figure A.5: Observed Sample versus Bootstrapped Sample



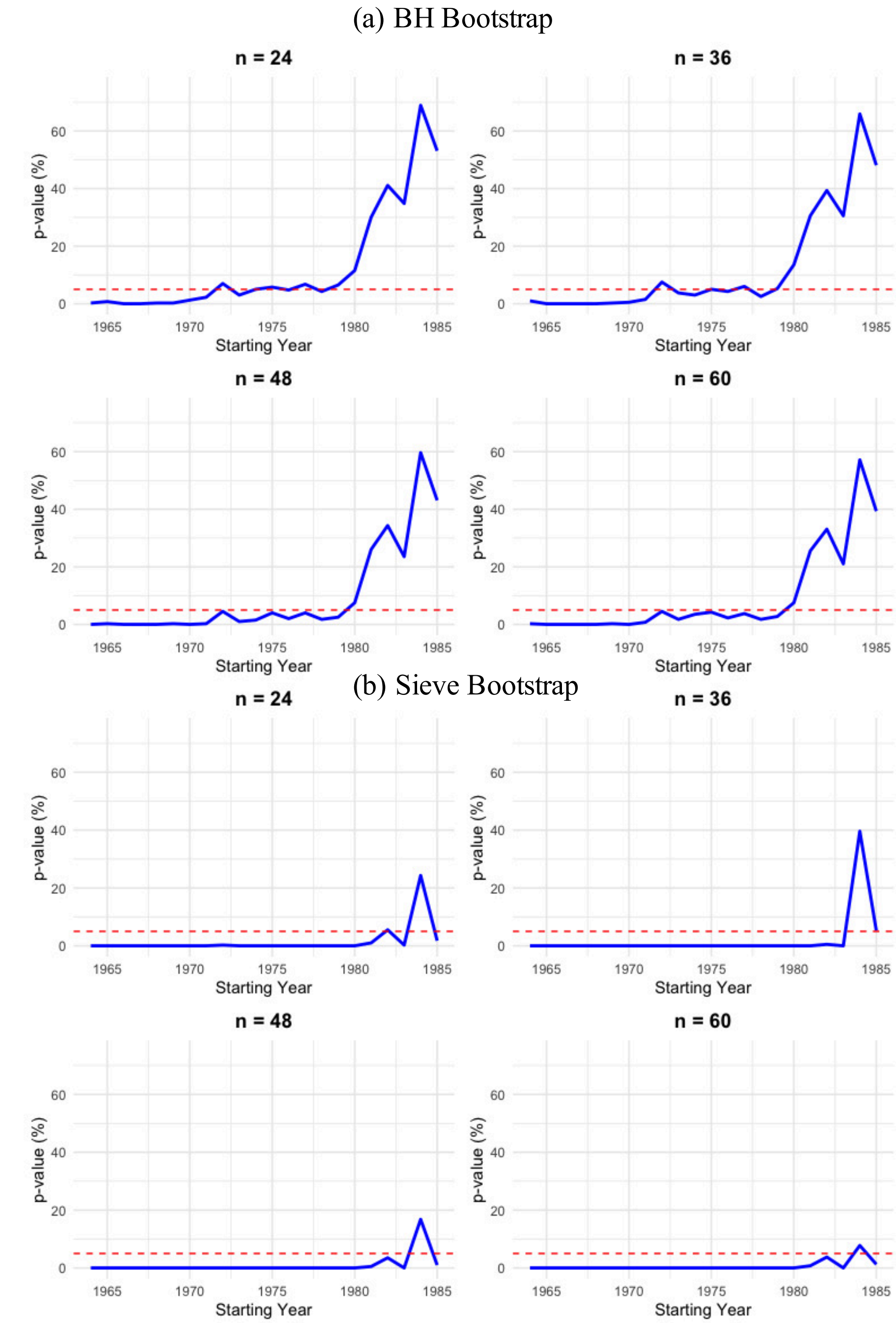
Notes: The left column presents the observed time series of yields, forward rates, and excess returns based on the Fama-Bliss discount bond data for the sample period 1964:Q1–2003:Q4. The right column displays the corresponding sieve bootstrapped time series for the same variables.

Figure A.6: Rolling R^2 Analysis Across Maturities



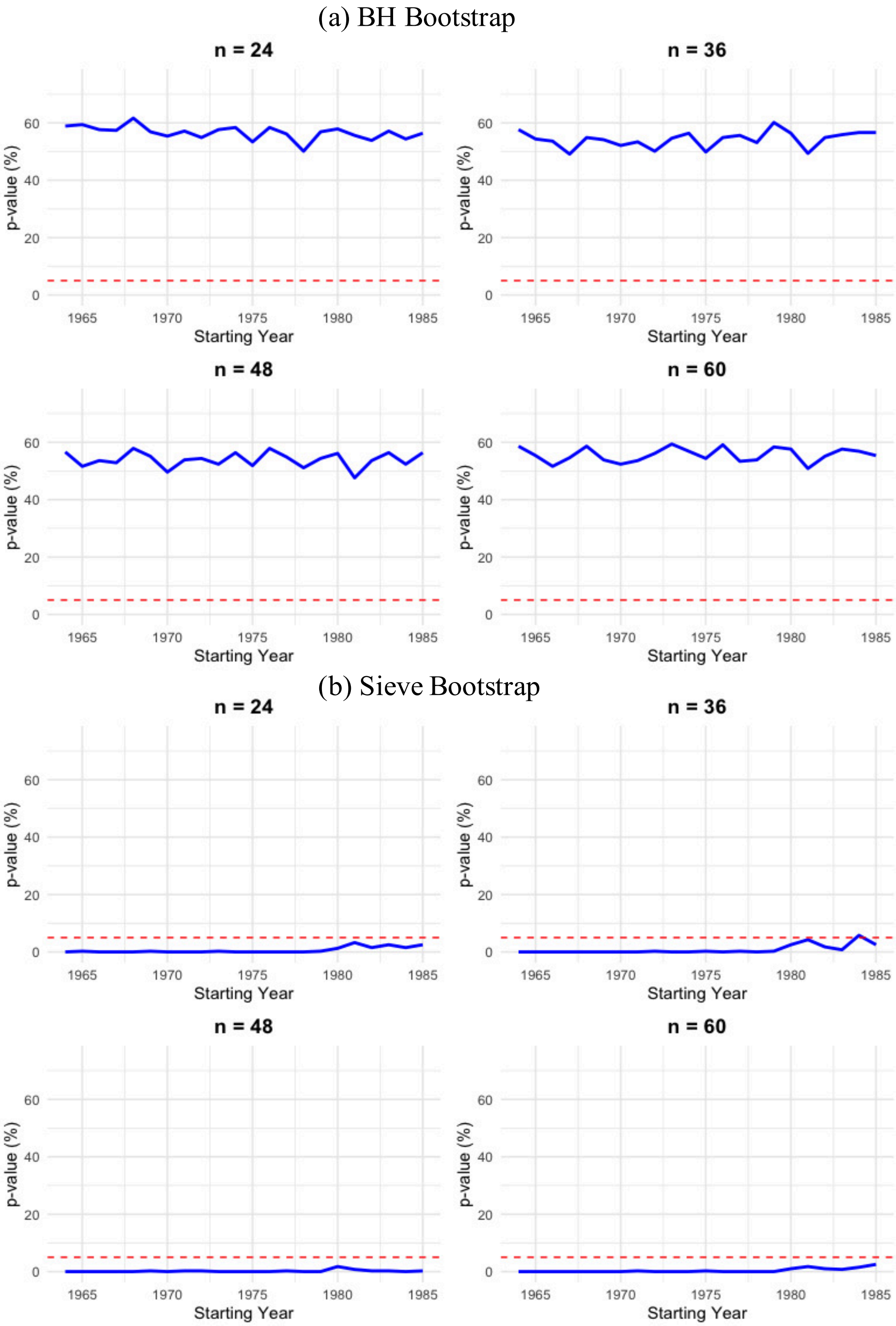
Notes: This plot illustrates the rolling average R^2 values for regressions of excess returns on the first three principal components (PC1-3) and CP's return-forecasting factor across four maturities $n \in \{24, 36, 48, 60\}$. The red line represents the regressions that include both PC1-3 and CP's factor. The blue line represents the regressions that include only PC1-3. The green line represents the regressions that include only the CP's factor and a constant. The analysis covers the period from 1964 to 2024, using a rolling window of 39 years.

Figure A.7: Rolling Bootstrapped p-values for Studentized Tests



Notes: The top row presents rolling studentized p-values from the BH bootstrap method, while the bottom row presents the corresponding results from the sieve bootstrap. The analysis spans the period 1964 to 2024, using a rolling window of 39 years and based on $B = 399$ bootstrap replications. The red dashed line indicates the 5% nominal level, and the p-values are expressed in percentage units.

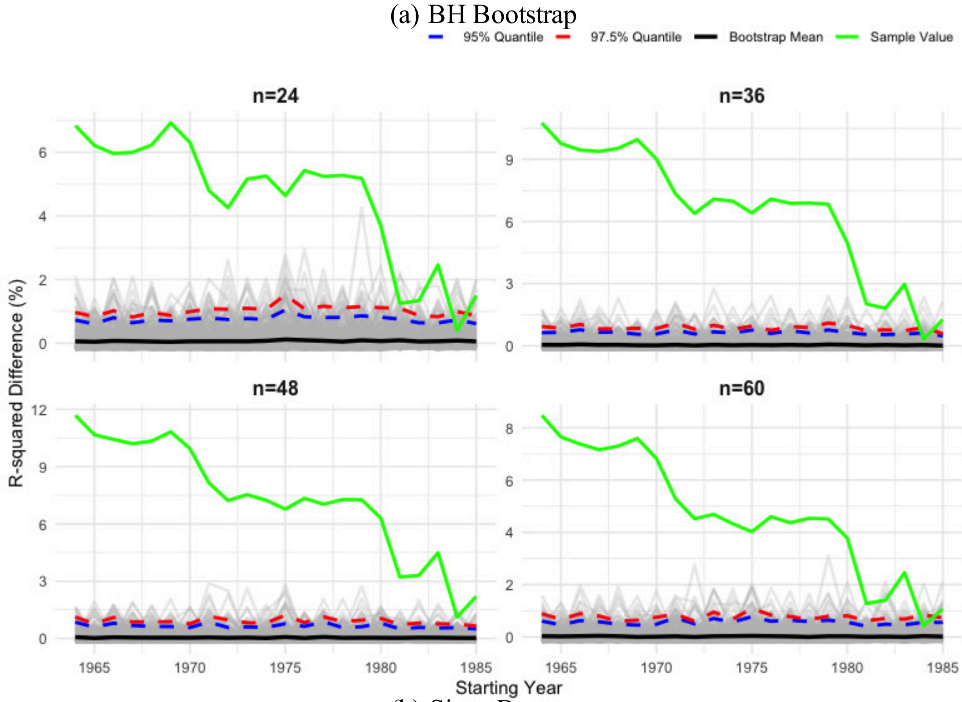
Figure A.8: Rolling Bootstrapped p-values for Unstudentized Tests



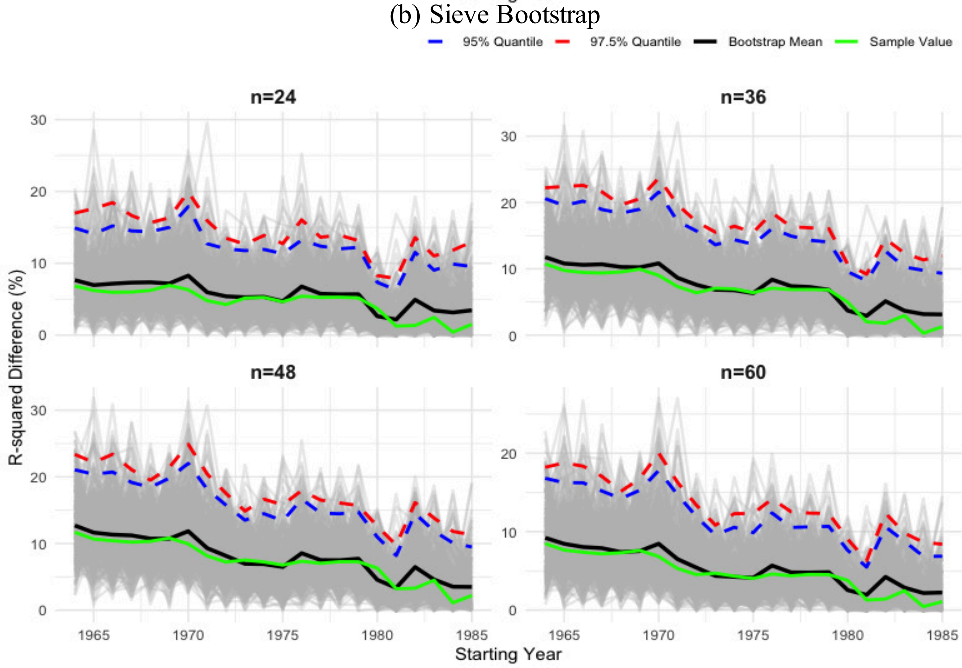
Notes: The top row presents rolling unstudentized p-values from the BH bootstrap method, while the bottom row presents the corresponding results from the sieve bootstrap. The analysis spans the period 1964 to 2024, using a rolling window of 39 years and based on $B = 399$ bootstrap replications. The red dashed line indicates the 5% nominal level, and the p-values are expressed in percentage units.

Figure A.9: Rolling Adjusted R^2 Differences Across Maturities

(a) BH Bootstrap

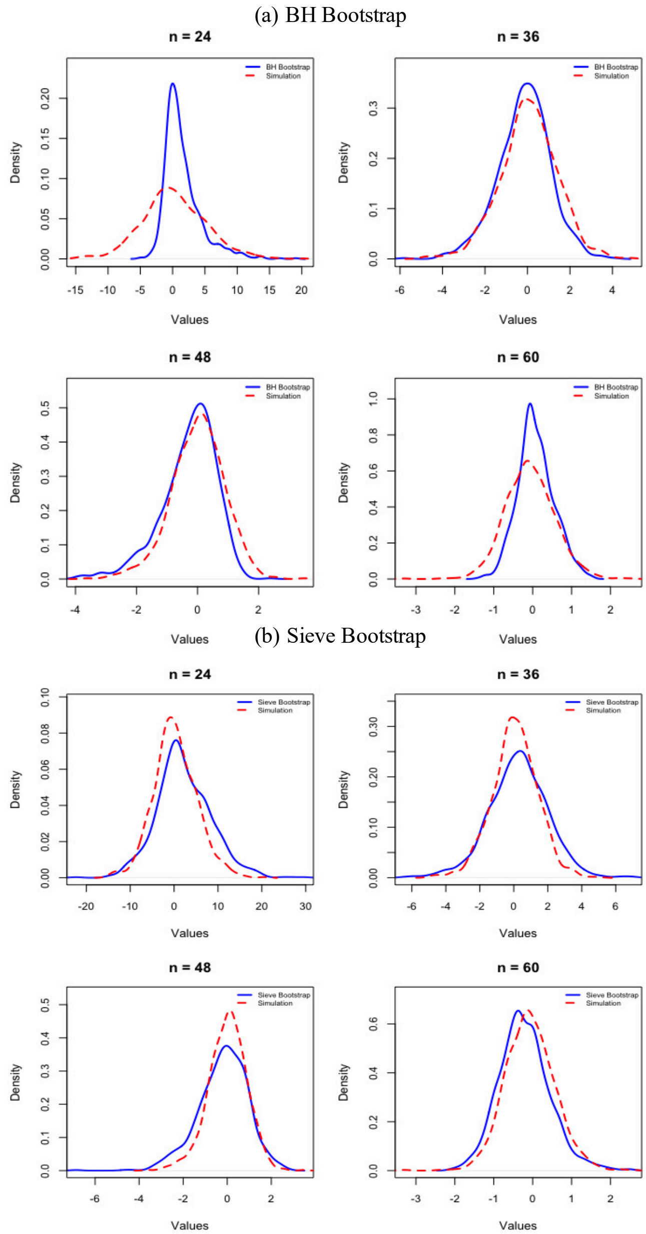


(b) Sieve Bootstrap



Notes: These plots illustrate the rolling differences in adjusted R^2 values across maturities $n \in \{24, 36, 48, 60\}$ for the BH and sieve bootstrap methods, respectively. The black line represents the bootstrap mean of the adjusted R^2 differences, while the green line shows the sample value differences. The dashed blue and red lines indicate the 95% and 97.5% quantiles of the bootstrap distribution, respectively. The shaded gray area represents the full range of bootstrap samples. The analysis spans the period 1964 to 2024, using a rolling window of 39 years and based on $B = 399$ bootstrap replications.

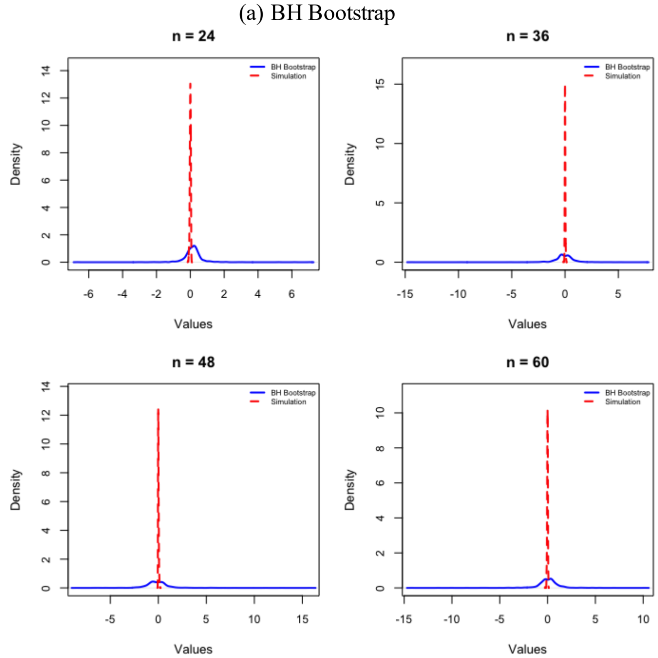
Figure A.10: Bootstrap Distributions Across Maturities under Studentized Tests



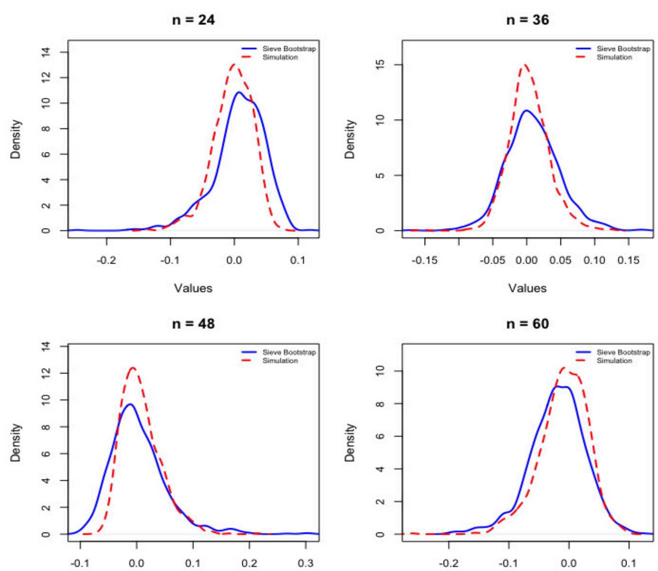
Notes: The figure illustrates the distributions of the bootstrapped estimates $\hat{\beta}^*_{2n,s(b)}$ across maturities $n \in \{24, 36, 48, 60\}$ under studentized tests. The top panel presents the distributions obtained using the BH bootstrap method, while the bottom panel shows the corresponding distributions derived from the sieve bootstrap method. The dashed red lines depict the distributions from the simulated sample, as described at the end of subsection 3.1, whereas the solid blue lines represent the empirical bootstrap distributions. Both bootstrap methods rely on the same simulated sample, with 1,000 bootstrap replications. For the sieve bootstrap method, the lag order is set to $p = 20$.

Figure A.11: Bootstrap Distributions Across Maturities under Unstudentized Tests

(a) BH Bootstrap



(b) Sieve Bootstrap



Notes: The figure depicts the distributions of the bootstrapped estimates $\hat{\beta}^*_{2n,s(b)}$ for maturities $n \in \{24, 36, 48, 60\}$ under unstudentized tests. The top panel illustrates the results obtained using the BH bootstrap method, while the bottom panel shows those generated by the sieve bootstrap method. The dashed red lines correspond to the distributions from the simulated sample, as outlined at the end of subsection 3.1, while the solid blue lines represent the bootstrap distributions based on the empirical data. Both bootstrap procedures utilize the same simulated sample, with 1,000 replications. For the sieve bootstrap, the lag order is set at $p = 20$.

Table A.1: Population Parameters for DGP in Section 5

Parameter	**Dimension**	**Values**
μ_f	1×1	0.14851
ρ_f	1×1	0.980026
$A_{1,12}$	5×5	$\begin{matrix} -0.5179 & 0.2617 & -0.0215 & 0.3684 & -0.1798 \\ -0.8640 & 0.3834 & 0.2681 & 0.2107 & -0.0849 \\ 0.0912 & -1.0020 & 0.4395 & 0.4532 & -0.0882 \\ 0.0754 & -0.0151 & -0.9245 & 0.7920 & 0.2115 \\ -0.0486 & -0.1334 & -0.1320 & -0.6271 & 1.0467 \end{matrix}$
$E[Z_t]$	1×5	$1.3943 \times 10^{-18}, \quad 0.0214, \quad -0.1947, \quad -0.2975, \quad -0.4476$
γ_n	5×4	$\begin{matrix} -0.9978 & -1.7303 & -2.4512 & -2.9902 \\ 0.7031 & 0.5823 & 0.8599 & 1.0906 \\ 1.1233 & 3.0541 & 3.7149 & 4.3816 \\ 0.2323 & 0.2969 & 1.1545 & 1.1058 \\ -0.9198 & -1.9197 & -2.8336 & -2.9633 \end{matrix}$
Σ_η	5×5	$\begin{matrix} 0.2241 & 0.0654 & 0.1058 & 0.1180 & 0.1101 \\ 0.0654 & 0.1936 & 0.1190 & 0.1296 & 0.1080 \\ 0.1058 & 0.1190 & 0.1847 & 0.1478 & 0.1529 \\ 0.1180 & 0.1296 & 0.1478 & 0.1928 & 0.1653 \\ 0.1101 & 0.1080 & 0.1529 & 0.1653 & 0.2503 \end{matrix}$

Table A.2: Size: Studentized t-Statistics at the 10% Nominal Level.

	Sieve Bootstrap					**CG Bootstrap**					**BH Bootstrap**				
Maturity	$Y_{12,t}$	$f_{24,12,t}$	$f_{36,12,t}$	$f_{48,12,t}$	$f_{60,12,t}$	$Y_{12,t}$	$f_{24,12,t}$	$f_{36,12,t}$	$f_{48,12,t}$	$f_{60,12,t}$	$Y_{12,t}$	$f_{24,12,t}$	$f_{36,12,t}$	$f_{48,12,t}$	$f_{60,12,t}$
2y	0.092	0.076	0.078	0.095	0.067	0.018	0.083	0.053	0.023	0.002	0.035	0.007	0.057	0.009	0.016
3y	0.077	0.084	0.071	0.077	0.096	0.003	0.017	0.067	0.054	0.021	0.003	0.022	0.005	0.044	0.008
4y	0.091	0.076	0.084	0.070	0.077	0.017	0.002	0.014	0.061	0.055	0.011	0.002	0.014	0.005	0.047
5y	0.080	0.097	0.076	0.092	0.071	0.059	0.020	0.001	0.012	0.046	0.063	0.009	0.009	0.014	0.003

Notes: This table presents the size results for three bootstrap methods with a sample size of $T = 800$. For the moving block bootstrap method proposed by Crump and Gospodinov (2024) (CG), a block size of $M = 28$ is used. For the sieve bootstrap, a VAR order of $p = 12$ is employed for the bootstrapping procedure. Each column provides results for the t-test, while each row corresponds to bond returns at a specific maturity.

Table A.3: Power: Studentized t-Statistics at the 10% Nominal Level.

	Sieve Bootstrap					**CG Bootstrap**					**BH Bootstrap**				
Maturity	$Y_{12,t}$	$f_{24,12,t}$	$f_{36,12,t}$	$f_{48,12,t}$	$f_{60,12,t}$	$Y_{12,t}$	$f_{24,12,t}$	$f_{36,12,t}$	$f_{48,12,t}$	$f_{60,12,t}$	$Y_{12,t}$	$f_{24,12,t}$	$f_{36,12,t}$	$f_{48,12,t}$	$f_{60,12,t}$
2y	0.990	0.976	0.108	1.000	0.220	0.736	0.974	0.084	0.866	0.016	0.694	0.016	0.018	0.347	0.347
3y	0.486	0.989	0.992	0.567	0.997	0.365	0.711	0.989	0.576	0.803	0.425	0.606	0.035	0.426	0.426
4y	0.961	0.201	0.991	0.996	0.407	0.425	0.055	0.739	0.991	0.384	0.071	0.056	0.572	0.055	0.055
5y	0.169	1.000	0.212	0.992	0.974	0.128	0.909	0.025	0.709	0.903	0.026	0.443	0.029	0.521	0.521

Notes: This table presents the power results for three bootstrap methods with a sample size of $T = 800$. For the moving block bootstrap method proposed by Crump and Gospodinov (2024) (CG), a block size of $M = 28$ is used. For the sieve bootstrap, a VAR order of $p = 12$ is employed for the bootstrapping procedure. Each column provides results for the t-test, while each row corresponds to bond returns at a specific maturity.

Table A.4: Size: Unstudentized t-Statistics at the 10% Nominal Level.

	Sieve Bootstrap					**CG Bootstrap**					**BH Bootstrap**				
Maturity	$Y_{12,t}$	$f_{24,12,t}$	$f_{36,12,t}$	$f_{48,12,t}$	$f_{60,12,t}$	$Y_{12,t}$	$f_{24,12,t}$	$f_{36,12,t}$	$f_{48,12,t}$	$f_{60,12,t}$	$Y_{12,t}$	$f_{24,12,t}$	$f_{36,12,t}$	$f_{48,12,t}$	$f_{60,12,t}$
2y	0.105	0.091	0.079	0.100	0.081	0.039	0.053	0.047	0.046	0.077	0.044	0.005	0.046	0.002	0.036
3y	0.098	0.101	0.096	0.077	0.102	0.067	0.033	0.052	0.051	0.040	0.002	0.034	0.002	0.024	0.002
4y	0.106	0.095	0.090	0.097	0.079	0.043	0.060	0.022	0.044	0.053	0.003	0.002	0.019	0.001	0.022
5y	0.089	0.103	0.086	0.088	0.093	0.046	0.042	0.059	0.022	0.033	0.060	0.002	0.023	0.014	0.001

Notes: This table presents the size results for three bootstrap methods with a sample size of $T = 800$. For the moving block bootstrap method proposed by Crump and Gospodinov (2024) (CG), a block size of $M = 28$ is used. For the sieve bootstrap, a VAR order of $p = 12$ is employed for the bootstrapping procedure. Each column provides results for the t-test, while each row corresponds to bond returns at a specific maturity.

Table A.5: Power: Unstudentized t-Statistics at the 10% Nominal Level.

	Sieve Bootstrap					**CG Bootstrap**					**BH Bootstrap**				
Maturity	$Y_{12,t}$	$f_{24,12,t}$	$f_{36,12,t}$	$f_{48,12,t}$	$f_{60,12,t}$	$Y_{12,t}$	$f_{24,12,t}$	$f_{36,12,t}$	$f_{48,12,t}$	$f_{60,12,t}$	$Y_{12,t}$	$f_{24,12,t}$	$f_{36,12,t}$	$f_{48,12,t}$	$f_{60,12,t}$
2y	0.996	0.993	0.142	1.000	0.247	0.991	0.987	0.101	1.000	0.236	0.894	0.003	0.003	0.119	0.002
3y	0.577	0.994	1.000	0.630	1.000	0.647	0.990	0.997	0.639	1.000	0.611	0.810	0.006	0.498	0.008
4y	0.987	0.228	1.000	1.000	0.462	0.963	0.326	0.991	0.999	0.455	0.000	0.067	0.713	0.006	0.170
5y	0.214	1.000	0.249	0.997	0.991	0.140	1.000	0.287	0.989	0.971	0.008	0.784	0.008	0.609	0.002

Notes: This table presents the power results for three bootstrap methods with a sample size of $T = 800$. For the moving block bootstrap method proposed by Crump and Gospodinov (2024) (CG), a block size of $M = 28$ is used. For the sieve bootstrap, a VAR order of $p = 12$ is employed for the bootstrapping procedure. Each column provides results for the t-test, while each row corresponds to bond returns at a specific maturity.

Table A.6: Size: Studentized t-Statistics at the 5% Nominal Level.

	Sieve Bootstrap					**CG Bootstrap**					**BH Bootstrap**				
Maturity	$Y_{12,t}$	$f_{24,12,t}$	$f_{36,12,t}$	$f_{48,12,t}$	$f_{60,12,t}$	$Y_{12,t}$	$f_{24,12,t}$	$f_{36,12,t}$	$f_{48,12,t}$	$f_{60,12,t}$	$Y_{12,t}$	$f_{24,12,t}$	$f_{36,12,t}$	$f_{48,12,t}$	$f_{60,12,t}$
2y	0.052	0.029	0.037	0.046	0.028	0.010	0.050	0.026	0.009	0.000	0.020	0.006	0.030	0.006	0.005
3y	0.034	0.046	0.029	0.038	0.048	0.000	0.009	0.043	0.024	0.008	0.001	0.015	0.004	0.024	0.006
4y	0.045	0.036	0.045	0.028	0.036	0.011	0.000	0.004	0.037	0.023	0.006	0.002	0.010	0.003	0.025
5y	0.036	0.047	0.033	0.041	0.029	0.027	0.008	0.000	0.005	0.028	0.033	0.005	0.002	0.008	0.001

Notes: This table presents the size results for three bootstrap methods with a sample size of $T = 800$. For the moving block bootstrap method proposed by Crump and Gospodinov (2024) (CG), a block size of $M = 28$ is used. For the sieve bootstrap, a VAR order of $p = 12$ is employed for the bootstrapping procedure. Each column provides results for the t-test, while each row corresponds to bond returns at a specific maturity.

Table A.7: Power: Studentized t-Statistics at the 5% Nominal Level.

	Sieve Bootstrap					**CG Bootstrap**					**BH Bootstrap**				
Maturity	$Y_{12,t}$	$f_{24,12,t}$	$f_{36,12,t}$	$f_{48,12,t}$	$f_{60,12,t}$	$Y_{12,t}$	$f_{24,12,t}$	$f_{36,12,t}$	$f_{48,12,t}$	$f_{60,12,t}$	$Y_{12,t}$	$f_{24,12,t}$	$f_{36,12,t}$	$f_{48,12,t}$	$f_{60,12,t}$
2y	0.964	0.922	0.052	0.992	0.134	0.528	0.914	0.042	0.644	0.006	0.545	0.008	0.004	0.215	0.002
3y	0.343	0.960	0.969	0.406	0.990	0.210	0.518	0.934	0.405	0.530	0.265	0.451	0.007	0.262	0.152
4y	0.900	0.117	0.978	0.988	0.258	0.204	0.020	0.541	0.946	0.250	0.036	0.024	0.414	0.027	0.101
5y	0.083	0.997	0.132	0.974	0.912	0.060	0.754	0.011	0.532	0.734	0.010	0.308	0.008	0.381	0.002

Notes: This table presents the power results for three bootstrap methods with a sample size of $T = 800$. For the moving block bootstrap method proposed by Crump and Gospodinov (2024) (CG), a block size of $M = 28$ is used. For the sieve bootstrap, a VAR order of $p = 12$ is employed for the bootstrapping procedure. Each column provides results for the t-test, while each row corresponds to bond returns at a specific maturity.

Table A.8: Size: Unstudentized t-Statistics at the 5% Nominal Level.

	Sieve Bootstrap					**CG Bootstrap**					**BH Bootstrap**				
Maturity	$Y_{12,t}$	$f_{24,12,t}$	$f_{36,12,t}$	$f_{48,12,t}$	$f_{60,12,t}$	$Y_{12,t}$	$f_{24,12,t}$	$f_{36,12,t}$	$f_{48,12,t}$	$f_{60,12,t}$	$Y_{12,t}$	$f_{24,12,t}$	$f_{36,12,t}$	$f_{48,12,t}$	$f_{60,12,t}$
2y	0.055	0.051	0.045	0.054	0.035	0.020	0.028	0.023	0.024	0.031	0.026	0.003	0.022	0.001	0.012
3y	0.049	0.050	0.056	0.042	0.054	0.019	0.019	0.023	0.021	0.027	0.000	0.022	0.002	0.006	0.002
4y	0.042	0.044	0.045	0.051	0.043	0.017	0.022	0.013	0.021	0.027	0.001	0.000	0.009	0.001	0.006
5y	0.047	0.051	0.041	0.044	0.051	0.027	0.022	0.023	0.013	0.015	0.025	0.001	0.005	0.006	0.001

Notes: This table presents the size results for three bootstrap methods with a sample size of $T = 800$. For the moving block bootstrap method proposed by Crump and Gospodinov (2024) (CG), a block size of $M = 28$ is used. For the sieve bootstrap, a VAR order of $p = 12$ is employed for the bootstrapping procedure. Each column provides results for the t-test, while each row corresponds to bond returns at a specific maturity.

Table A.9: Power: Unstudentized t-Statistics at the 5% Nominal Level.

	Sieve Bootstrap					**CG Bootstrap**					**BH Bootstrap**				
Maturity	$Y_{12,t}$	$f_{24,12,t}$	$f_{36,12,t}$	$f_{48,12,t}$	$f_{60,12,t}$	$Y_{12,t}$	$f_{24,12,t}$	$f_{36,12,t}$	$f_{48,12,t}$	$f_{60,12,t}$	$Y_{12,t}$	$f_{24,12,t}$	$f_{36,12,t}$	$f_{48,12,t}$	$f_{60,12,t}$
2y	0.993	0.980	0.073	1.000	0.167	0.955	0.958	0.034	1.000	0.125	0.811	0.001	0.001	0.020	0.000
3y	0.457	0.991	0.996	0.508	0.999	0.576	0.947	0.994	0.510	0.999	0.490	0.677	0.001	0.273	0.001
4y	0.962	0.148	0.993	1.000	0.339	0.864	0.209	0.959	0.998	0.339	0.000	0.011	0.566	0.002	0.059
5y	0.127	1.000	0.169	0.993	0.985	0.054	1.000	0.159	0.956	0.915	0.002	0.396	0.001	0.456	0.000

Notes: This table presents the power results for three bootstrap methods with a sample size of $T = 800$. For the moving block bootstrap method proposed by Crump and Gospodinov (2024) (CG), a block size of $M = 28$ is used. For the sieve bootstrap, a VAR order of $p = 12$ is employed for the bootstrapping procedure. Each column provides results for the t-test, while each row corresponds to bond returns at a specific maturity.

Table A.10: CG Bootstrap Results for Different Moving Sizes ($M = 28, M = 75$), 10% Nominal Level.

	Size ($M = 28$)					**Size** ($M = 75$)				
Maturity	$Y_{12,t}$	$f_{24,12,t}$	$f_{36,12,t}$	$f_{48,12,t}$	$f_{60,12,t}$	$Y_{12,t}$	$f_{24,12,t}$	$f_{36,12,t}$	$f_{48,12,t}$	$f_{60,12,t}$
2y	0.018	0.083	0.053	0.023	0.002	0.061	0.102	0.077	0.061	0.044
3y	0.003	0.017	0.067	0.054	0.021	0.055	0.058	0.095	0.075	0.060
4y	0.017	0.002	0.014	0.061	0.055	0.062	0.050	0.049	0.093	0.072
5y	0.059	0.020	0.001	0.012	0.046	0.087	0.060	0.044	0.048	0.082

Notes: This table reports the size results for two moving block sizes $M = 28$ and $M = 75$. The moving block bootstrap method is based on Crump and Gospodinov (2024) (CG). Each column provides results for the t-test, while each row corresponds to bond returns at a specific maturity.

Table A.11: CG Bootstrap Results for Different Moving Sizes ($M = 28, M = 75$), 10% Nominal Level.

	Power ($M = 28$)					**Power** ($M = 75$)				
Maturity	$Y_{12,t}$	$f_{24,12,t}$	$f_{36,12,t}$	$f_{48,12,t}$	$f_{60,12,t}$	$Y_{12,t}$	$f_{24,12,t}$	$f_{36,12,t}$	$f_{48,12,t}$	$f_{60,12,t}$
2y	0.736	0.974	0.084	0.866	0.016	0.961	0.974	0.125	0.985	0.134
3y	0.365	0.711	0.989	0.576	0.803	0.470	0.956	0.990	0.558	0.980
4y	0.425	0.055	0.739	0.991	0.384	0.921	0.157	0.964	0.989	0.380
5y	0.128	0.909	0.025	0.709	0.903	0.186	0.990	0.150	0.958	0.957

Notes: This table reports the power results for two moving block sizes $M = 28$ and $M = 75$. The moving block bootstrap method is based on Crump and Gospodinov (2024) (CG). Each column provides results for the t-test, while each row corresponds to bond returns at a specific maturity.

Table A.12: Sieve Bootstrap Results for Different Lags ($p = 12, p = 15, p = 20$), 10% Nominal Level.

	Size ($p = 12$)					**Size** ($p = 15$)					**Size** ($p = 20$)				
Maturity	$Y_{12,t}$	$f_{24,12,t}$	$f_{36,12,t}$	$f_{48,12,t}$	$f_{60,12,t}$	$Y_{12,t}$	$f_{24,12,t}$	$f_{36,12,t}$	$f_{48,12,t}$	$f_{60,12,t}$	$Y_{12,t}$	$f_{24,12,t}$	$f_{36,12,t}$	$f_{48,12,t}$	$f_{60,12,t}$
2y	0.092	0.076	0.078	0.095	0.067	0.098	0.077	0.078	0.095	0.071	0.094	0.077	0.075	0.084	0.067
3y	0.077	0.084	0.071	0.077	0.096	0.070	0.088	0.070	0.078	0.092	0.073	0.086	0.070	0.076	0.089
4y	0.091	0.076	0.084	0.070	0.077	0.095	0.069	0.091	0.070	0.075	0.095	0.074	0.088	0.075	0.081
5y	0.080	0.097	0.076	0.092	0.071	0.082	0.095	0.070	0.092	0.070	0.080	0.092	0.073	0.093	0.076

Notes: This table reports the size results for three lag lengths $p = 12$, $p = 15$, and $p = 20$. Each column provides results for the t-test, while each row corresponds to bond returns at a specific maturity.

Table A.13: Statistical inference in excess return regressions

A. Original sample, 1964–2003

	n=24					n=36				
	PC1	$F_{1,t}$ PC2	PC3	$F_{2,t}$ CP factor	R^2	PC1	$F_{1,t}$ PC2	PC3	$F_{2,t}$ CP factor	R^2
Coefficient	0.0329	0.106	-0.689	0.398	0.257	-0.003	-0.259	-0.049	0.911	0.239
HAC p-value	**0.000**	0.457	0.622	**0.000**	0.320	0.576	0.578	0.991	**0.000**	0.338
BH p-value				**0.000**					**0.005**	
				0.617					0.559	
Sieve p-value				**0.000**					**0.000**	
				0.000					**0.000**	

	n=48					n=60				
	PC1	$F_{1,t}$ PC2	PC3	$F_{2,t}$ CP factor	R^2	PC1	$F_{1,t}$ PC2	PC3	$F_{2,t}$ CP factor	R^2
Coefficient	-0.024	-0.267	0.797	1.317	0.262	-0.006	0.420	-0.058	1.374	0.267
HAC p-value	0.238	0.750	0.915	**0.000**	0.371	0.722	0.738	0.996	**0.000**	0.346
BH p-value				**0.000**					**0.000**	
				0.536					0.564	
Sieve p-value				**0.000**					**0.000**	
				0.000					**0.000**	

B. Later sample, 1985–2019

	n=24					n=36				
	PC1	$F_{1,t}$ PC2	PC3	$F_{2,t}$ CP factor	R^2	PC1	$F_{1,t}$ PC2	PC3	$F_{2,t}$ CP factor	R^2
Coefficient	0.018	-0.218	-0.209	0.460	0.178	0.018	-0.156	-0.150	0.817	0.158
HAC p-value	**0.000**	0.486	0.910	**0.000**	0.204	**0.000**	0.898	0.984	**0.000**	0.180
BH p-value				0.409					0.383	
				0.539					0.539	
Sieve p-value				**0.023**					**0.028**	
				0.033					**0.040**	

	n=48					n=60				
	PC1	$F_{1,t}$ PC2	PC3	$F_{2,t}$ CP factor	R^2	PC1	$F_{1,t}$ PC2	PC3	$F_{2,t}$ CP factor	R^2
Coefficient	-0.029	-0.286	-0.895	1.432	0.171	-0.007	0.660	1.254	1.290	0.172
HAC p-value	**0.000**	0.900	0.952	**0.001**	0.204	0.648	0.850	0.959	0.075	0.189
BH p-value				0.321					0.281	
				0.549					0.599	
Sieve p-value				**0.008**					**0.023**	
				0.005					**0.040**	

Notes: This table presents the sample and bootstrapped regression results of annual excess bond returns on the first three principal components (PCs) of yields and CP's return-forecasting factor across four maturities. The null hypothesis tests whether the first three PCs encompass all relevant predictive information. Panel A data align with CP's original paper. HAC p-values are computed with Newey-West standard errors using 12 lags. Unadjusted R^2 values are reported for sample regressions with and without the CP factor. For the bootstrap methods, both studentized and unstudentized p-values are presented, where unstudentized p-values are listed in the second row of each method. p-values less than 5% are highlighted in bold.

References

[1] Andrews, D. W., & Monahan, J. C. (1992). An improved heteroskedasticity and autocorrelation consistent covariance matrix estimator. *Econometrica, 60*(4), 953–966.

[2] Bansal, R., & Shaliastovich, I. (2013). A long-run risks explanation of predictability puzzles in bond and currency markets. *The Review of Financial Studies, 26*(1), 1–33.

[3] Bauer, M. D., & Hamilton, J. D. (2018). Robust bond risk premia. *The Review of Financial Studies, 31*(2), 399–448.

[4] Brockwell, P. J., & Davis, R. A. (1987). *Time series: Theory and methods*. New York: Springer.

[5] Bühlmann, P. (1997). Sieve bootstrap for time series. *Annals of Statistics, 25*(3), 1045–1067.

[6] Campbell, J. Y., & Yogo, M. (2006). Efficient tests of stock return predictability. *Journal of Financial Economics, 81*(1), 27–60.

[7] Cavanagh, C. L., Elliott, G., & Stock, J. H. (1995). Inference in models with nearly integrated regressors. *Econometric Theory, 11*(6), 1131–1147.

[8] Cieslak, A., & Povala, P. (2015). Expected returns in Treasury bonds. *The Review of Financial Studies, 28*(10), 2859–2901.

[9] Cochrane, J. H., & Piazzesi, M. (2005). Bond risk premia. *American Economic Review, 95*(1), 138–160.

[10] Cochrane, J. H., & Piazzesi, M. (2009). Decomposing the yield curve. In *AFA 2010 Atlanta Meetings Paper*.

[11] Cooper, I., & Priestley, R. (2008). Time-varying risk premiums and the output gap. *The Review of Financial Studies, 22*(7), 2801–2833.

[12] Crump, R. K., & Gospodinov, N. (2022). On the factor structure of bond returns. *Econometrica, 90*(1), 295–314.

[13] Crump, R. K., & Gospodinov, N. (2024). Deconstructing the yield curve. *The Review of Financial Studies*. https://doi.org/10.1093/rfs/hhae077

[14] Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge: Cambridge University Press.

[15] Duffee, G. R. (2013). Bond pricing and the macroeconomy. In M. Harris, G. M. Constantinides, & R. M. Stulz (Eds.), *Handbook of the Economics of Finance* (Vol. 2, pp. 907–967). Elsevier.

[16] Goetzmann, W. N., & Jorion, P. (1993). Testing the predictive power of dividend yields. *The Journal of Finance, 48*(2), 663–679.

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

[18] Joslin, S., Priebsch, M., & Singleton, K. J. (2014). Risk premiums in dynamic term structure models with unspanned macro risks. *The Journal of Finance, 69*(3), 1197–1233.

[19] Kilian, L. (1998). Small-sample confidence intervals for impulse response functions. *The Review of Economics and Statistics, 80*(2), 218–230.

[20] Litterman, R. B., Scheinkman, J., & Weiss, L. (1991). Volatility and the yield curve. *The Journal of Fixed Income, 1*(1), 49–53.

[21] Ludvigson, S. C., & Ng, S. (2009). Macro factors in bond risk premia. *The Review of Financial Studies, 22*(12), 5027–5067.

[22] Ludvigson, S. C., & Ng, S. (2010). A factor analysis of bond risk premia. In M. E. Thornton & J. Wei (Eds.), *Handbook of Empirical Economics and Finance* (pp. 313–339). CRC Press.

[23] Lütkepohl, H. (2005). *New introduction to multiple time series analysis*. Springer Science & Business Media.

[24] Meyer, M., & Kreiss, J. P. (2015). On the vector autoregressive sieve bootstrap. *Journal of Time Series Analysis, 36*(3), 377–397.

[25] Müller, U. K. (2014). HAC corrections for strongly autocorrelated time series. *Journal of Business and Economic Statistics, 32*(3), 311–322.

[26] Newey, W. K., & West, K. D. (1987). Hypothesis testing with efficient method of moments estimation. *International Economic Review, 28*(3), 777–787.

[27] Niebuhr, T., Kreiss, J. P., & Paparoditis, E. (2017). Some properties of the autoregressive-aided block bootstrap. *Scandinavian Journal of Statistics, 44*(3), 678–707.

[28] Phillips, P. C. B. (1988). Regression theory for near-integrated time series. *Econometrica, 56*(5), 1021–1043.

[29] Politis, D. N., & Romano, J. P. (1994). The stationary bootstrap. *Journal of the American Statistical Association, 89*(428), 1303–1313.

[30] Reinsel, G. C. (1993). *Multivariate time series analysis*. New York: John Wiley & Sons.

[31] Stambaugh, R. F. (1999). Predictive regressions. *Journal of Financial Economics, 54*(3), 375–421.

[32] Sun, Y., & Kaplan, D. (2010). A new asymptotic theory for vector autoregressive long-run variance estimation and autocorrelation robust testing. Available at SSRN: https://doi.org/10.2139/ssrn.1714711.

[33] Valkanov, R. (2003). Long-horizon regressions: theoretical results and applications. *Journal of Financial Economics, 68*(2), 201–232.

[34] Wei, M., & Wright, J. H. (2013). Reverse regressions and long-horizon forecasting. *Journal of Applied Econometrics, 28*(3), 353–371.

[35] Xu, K. L. (2020). Inference of local regression in the presence of nuisance parameters. *Journal of Econometrics, 218*(2), 532–560.