FEDERAL RESERVE BANK OF SAN FRANCISCO

WORKING PAPER SERIES

Robust Bond Risk Premia

Michael D. Bauer
Federal Reserve Bank of San Francisco

James D. Hamilton
University of California, San Diego

May 2017

Working Paper 2015-15
http://www.frbsf.org/economic-research/publications/working-papers/wp2015-15.pdf

Suggested citation:

Bauer, Michael D., James D. Hamilton. 2017. "Robust Bond Risk Premia." Federal Reserve Bank of San Francisco Working Paper 2015-15. http://www.frbsf.org/economic-research/publications/working-papers/wp2015-15.pdf

Robust Bond Risk Premia*

Michael D. Bauer† and James D. Hamilton‡

April 16, 2015
Revised: May 22, 2017

Abstract

A consensus has recently emerged that variables beyond the level, slope, and curvature of the yield curve can help predict bond returns. This paper shows that the statistical tests underlying this evidence are subject to serious small-sample distortions. We propose more robust tests, including a novel bootstrap procedure specifically designed to test the spanning hypothesis. We revisit the analysis in six published studies and find that the evidence against the spanning hypothesis is much weaker than it originally appeared. Our results pose a serious challenge to the prevailing consensus.

Keywords: yield curve, spanning, return predictability, robust inference, bootstrap
JEL Classifications: E43, E44, E47

*The views expressed in this paper are those of the authors and do not necessarily reflect those of others in the Federal Reserve System. We thank Anna Cieslak, John Cochrane, Greg Duffee, Graham Elliott, Robin Greenwood, Helmut Lütkepohl, Ulrich Müller, Hashem Pesaran and Glenn Rudebusch for useful suggestions, conference participants and discussants at the 7th Annual Volatility Institute Conference at the NYU Stern School of Business, the NBER Summer Institute 2015, the Federal Reserve System Macro Conference 2015 in Cleveland, the Federal Reserve Bank of San Francisco Fixed Income Research Conference 2015, the CESifo Conference on Macro, Money and International Finance 2016 in Munich, the Spring 2016 NBER Asset Pricing Workshop in Chicago, and the Western Finance Association Conference 2016 in Park City, as well as seminar participants at the Federal Reserve Bank of Boston, the Free University of Berlin, and the University of Hamburg for helpful comments, and Anh Le, Marcel Priebsch, Serena Ng, Robin Greenwood, Richard Priestley and Anna Cieslak for the data used in their papers.

†Federal Reserve Bank of San Francisco, 101 Market St MS 1130, San Francisco, CA 94105, phone: 415-974-3299, e-mail: michael.bauer@sf.frb.org

‡University of California at San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0508, phone: 858-534-5986, e-mail: jhamilton@ucsd.edu

1 Introduction

Identifying the contribution of risk premia to long-term interest rates is crucial for monetary policy, investment strategy, and interpreting historical episodes such as the unprecedented low interest rates since 2008. Since the risk premium is just the difference between the current long rate and the expected average value of future short rates, the core question for estimating risk premia is how to construct short-rate expectations. Is it sufficient to consider the current yield curve, or should estimates incorporate additional information such as macroeconomic variables? This is the question we address in this paper.

A powerful theoretical argument suggests that the current yield curve itself should contain most (if not all) information useful for forecasting future interest rates and bond returns. Investors use information at time t—which we can summarize by a state vector z_t—to forecast future interest rates and risk premia. The price of a zero-coupon bond is thus a function of z_t and its maturity. The yield curve results from the prices of bonds with many different maturities, each of which is a different function of z_t. Under quite general assumptions the yield curve therefore contains the same information as z_t, since z_t can be inferred from yields. In other words, the yield curve spans all information relevant for forecasting future yields and returns, and no variables other than the current yield curve are needed. While this "spanning hypothesis" could be violated for various reasons, it is in fact implied by essentially all macro-finance models.[1] Therefore the spanning hypothesis is the natural benchmark when investigating the empirical relevance of macroeconomic and other variables for predictions of excess bond returns and estimation of bond risk premia, and a large literature has taken it as the relevant null hypothesis. Recent literature reviews by Gürkaynak and Wright (2012) and Duffee (2013a) identify the spanning hypothesis as a central issue in macro-finance. If it holds true it would greatly simplify forecasting of interest rates and estimation of monetary policy expectations and bond risk premia, as such forecasts and estimates would not require any macroeconomic series, other asset prices or quantities, volatilities, or survey expectations, but only the information in the current yield curve.[2]

Importantly, the spanning hypothesis does not imply that macroeconomic variables are unimportant for interest rates and risk premia. Quite to the contrary, interest rates are of course driven by macro variables in many ways, an obvious example being the importance of inflation expectations for nominal yields.[3] The yield curve reflects the information in current

[1] Examples of equilibrium models of the term structure that imply spanning include Wachter (2006), Piazzesi and Schneider (2007), Bekaert et al. (2009), and Bansal and Shaliastovich (2013). Macro-finance models with production economies (i.e., DSGE models) that imply spanning include Hördahl et al. (2006), Dewachter and Lyrio (2006), Rudebusch and Wu (2008), and Rudebusch and Swanson (2012).

[2] We will discuss the spanning hypothesis and its theoretical underpinnings in more detail in Section 2.1.

[3] Much theoretical and empirical work has investigated the links between macroeconomic variables, interest

and future macro variables, and the spanning hypothesis simply posits that it *fully* reflects and spans this information. Macroeconomic variables are drivers of risk premia, but our question here is what variables should be used for the *estimation* of these risk premia.

How should we summarize the information in the yield curve to empirically test the spanning hypothesis? It has long been recognized that the first three principal components (PCs) of yields, commonly labeled level, slope, and curvature, provide an excellent empirical summary of the entire yield curve (Litterman and Scheinkman, 1991), as they explain almost all of the cross-sectional variance of observed yields. This motivates a specific version of the spanning hypothesis, a very practical and empirically focused interpretation of the question posed above: Do level, slope and curvature completely capture all the information that is useful for forecasting future yields and estimating bond risk premia? This is the question we focus on in this paper.[4]

There is a growing consensus in the literature that the spanning hypothesis can be rejected by the observed data. This evidence comes from predictive regressions for bond returns on various predictors, controlling for information in the current yield curve. The variables that have been found to contain additional predictive power in such regressions include measures of economic growth and inflation (Joslin et al., 2014), factors inferred from a large set of macro variables (Ludvigson and Ng, 2009, 2010), long-term trends in inflation or inflation expectations (Cieslak and Povala, 2015), the output gap (Cooper and Priestley, 2008), and measures of Treasury bond supply (Greenwood and Vayanos, 2014). These results suggest that there might be unspanned or hidden information that is not captured by the current yield curve but that is useful for forecasting. In addition, Cochrane and Piazzesi (2005) found that higher-order (fourth and fifth) PCs of bond yields appear useful for predicting bond returns, which suggests that the miniscule amount of yield variation not captured by the first three PCs somehow contains relevant information about bond risk premia.

But these predictive regressions have a number of problematic features. The true predictive variables under the null hypothesis are necessarily correlated with lagged forecast errors because they summarize the information in the current yield curve. As a consequence they violate the condition of strict econometric exogeneity. In addition, the predictive variables are typically highly persistent. We show that this leads to substantial "standard error bias" in samples of the size commonly studied, with the problem even more severe when the proposed explanatory variables exhibit a trend over the observed sample. Because the estimated standard errors are too small, the result can often be spurious rejection of the spanning hypothesis

rates, and risk premia. Some prominent examples include Campbell and Cochrane (1999), Diebold et al. (2006), Bikbov and Chernov (2010), Rudebusch and Swanson (2012), and Bansal and Shaliastovich (2013).

[4]While most of our analysis centers on this question, we also report results for alternative spanning hypotheses under which four or five PCs fully capture the information in the yield curve.

even though it is true. This problem inherent in all tests of the spanning hypothesis has to our knowledge not previously been recognized. Mankiw and Shapiro (1986) and Stambaugh (1999) documented small-sample coefficient bias in predictive regressions with a persistent regressor that is not strictly exogenous.[5] By contrast, in our setting there is no coefficient bias pertaining to the additional predictors, and instead a downward bias of the estimated standard errors distorts the results of conventional inference. An additional problem is that the common predictive regressions are estimated in monthly data but with an annual excess bond return as the dependent variable, and the presence of overlapping observations introduces substantial serial correlation in the prediction errors. As a result, standard errors are even less reliable, and regression R^2 are harder to interpret. We demonstrate that the procedures commonly used for inference about the spanning hypothesis do not adequately address these issues and are subject to serious small-sample distortions.

We propose a novel approach that researchers can use to obtain more robust inference in these predictive regressions: a parametric bootstrap that generates data samples under the spanning hypothesis. We calculate the first three PCs of the observed set of yields and summarize their dynamics with a VAR fit to the observed PCs. Then we use a residual bootstrap to resample the PCs, and construct bootstrapped yields by multiplying the simulated PCs by the historical loadings of yields on the PCs and adding a small Gaussian measurement error. Thus by construction no variables other than the PCs are useful for predicting yields or returns in our generated data. We then fit a separate VAR to the proposed additional explanatory variables alone, and generate bootstrap samples for the predictors from this VAR. Using our novel bootstrap procedure, we can calculate the properties of any regression statistic under the spanning hypothesis.[6] This calculation demonstrates that the conventional tests reject the true null much too often. We show that the tests employed in published studies, which are intended to have a nominal size of five percent, have a true size between 8 and 61%. We then use our bootstrap to ask how likely it would be under the null to observe the patterns of predictability that researchers have found in the data. We find that the proposed predictors are always much less significant than appeared in conventional tests, and are often statistically insignificant. These results provide a strong caution against using conventional tests for inference about bond risk premia, and we recommend that researchers instead use the bootstrap procedure proposed in this paper.

An additional way to assess the robustness of the published results is to take advantage

[5] Cavanagh et al. (1995) and Campbell and Yogo (2006) considered this problem using local-to-unity asymptotic theory.

[6] Our procedure notably differs from the bootstrap approach commonly employed in this literature, which generates artificial data under the expectations hypothesis, such as Bekaert et al. (1997), Cochrane and Piazzesi (2005), Ludvigson and Ng (2009, 2010), and Greenwood and Vayanos (2014).

of the data that have arrived since publication of these studies. In addition to re-estimating the proposed predictive models over a common, more recent data sample, we use the newly available data to evaluate whether they improve true out-of-sample forecasts, which gives us a more reliable test than the often-reported pseudo-out-of-sample statistics. We find that the proposed additional predictors are rarely helpful in the new data, reinforcing the case that the apparent strength of the in-sample evidence may be an artifact of the small-sample problems we highlight.

After revisiting the evidence in the six influential papers cited above we draw two main conclusions: First, conventional methods of inference are extremely unreliable in these predictive regressions, because they often suggest that variables are relevant for bond risk premia which in truth are irrelevant. New approaches for robust inference are needed, and we propose three in this paper. Second, when reconsidered with more robust methods for inference, the evidence against the spanning hypothesis appears weaker and much less robust than would appear from the published results, and in some cases appears to be spurious.

Our paper is related to other studies that critically assess return predictability in finance. For stock returns, Goetzmann and Jorion (1993) and Nelson and Kim (1993) used simulation studies to document that small-sample problems can lead to spurious findings of predictability. Ferson et al. (2003) raised the possibility of finding spurious predictability if a persistent component of stock returns is unobserved. Ang and Bekaert (2007) demonstrated that the commonly employed Newey-West standard errors are not reliable for long-horizon predictions. Welch and Goyal (2008) showed that predictability largely disappears in out-of-sample analysis, and Lewellen et al. (2010) showed that estimating factor models for equity risk premia can lead to spuriously high R^2 for truly irrelevant risk factors. Our paper parallels these studies by also documenting that published evidence on predictability and risk premia is fraught with serious econometric problems and appears to be partially spurious. But our work is distinct in that we describe a new, different econometric issue and focus on evidence on unspanned risks in bond returns instead of predictability of stock returns. The literature on bond returns and the expectations hypothesis goes back to Fama and Bliss (1987) and Campbell and Shiller (1991), who established that the slope of the yield curve helps predict bond returns. Bekaert et al. (1997) and Bekaert and Hodrick (2001) documented that rejections of the expectations hypothesis are robust to the Stambaugh bias that arises in predictive regressions for bond returns. Our paper shows that a different kind of bias—standard error bias—arises in the widely used tests of the spanning hypothesis, and that accounting for it can change the empirical conclusions.

2 Inference about the spanning hypothesis

In this section we first explain the economic underpinnings and common empirical tests of the spanning hypothesis, and then document previously unrecognized econometric problems with these tests. Then we propose a new way of inference about the spanning hypothesis that solves these problems using an easy-to-implement parametric bootstrap.

2.1 The spanning hypothesis

A simple but powerful theoretical argument demonstrates that under certain assumptions about financial markets the yield curve fully spans the information set that is relevant for forecasting future interest rates and estimating risk premia.[7] If the vector z_t denotes the information that investors use for pricing financial assets, then bond prices and yields are functions of z_t. Since bond yields are determined by investor's expectations of future short-term rates and future excess returns, z_t contains the information required to construct these forecasts. For example, z_t would likely contain macroeconomic variables that matter for interest rates, such as current and expected future inflation. Denoting by Y_t a vector of N yields of different maturities we have $Y_t = f(z_t)$ where $f(\cdot)$ is a vector-valued function. The spanning hypothesis assumes that f is invertible, in which case the information in z_t can be inferred as $z_t = f^{-1}(Y_t)$. A necessary condition for this invertibility condition is that N is at least as large as the number of variables in z_t, which is a plausible assumption given the large number of yields that constitute the yield curve. Invertibility is guaranteed for example if f is linear and its Jabobian has full column rank, but it will also hold under much more general conditions. Most asset pricing and macro-finance models imply invertibility of model-implied yields for state variables, hence the spanning hypothesis holds in these models.[8] As mentioned above, the spanning hypothesis of course does not imply that macro variables are unimportant for interest rates, but simply states that the yield curve fully spans the relevant information in macro (and other) variables.

While essentially all asset pricing models imply some version of spanning, there are a number of reasons why the relevant information may not be spanned by the first three PCs of observed yields, which is the null hypothesis we focus on in this paper. First, yields may of course depend on more than three state variables. For example, in Bansal and Shaliastovich (2013) yields are functions of four state variables. Second, even if three linear combinations of model-implied yields span z_t, this might be difficult to exploit in practice due to measurement error. In particular, Duffee (2011b) demonstrated that if the effects of some elements of z_t on

[7] This argument largely follows the one in Duffee (2013b, Section 2.3).

[8] See footnote 1 for relevant references on this point.

yields nearly offset each other, those components will be very difficult to infer from current observed yields alone. Cieslak and Povala (2015) and Bauer and Rudebusch (2017) noted that in affine yield-curve models, even small measurement errors can make it impossible to recover z_t from observed yields. Third, statistical expectations may differ from subjective expectations due to learning (as in, for example, Piazzesi et al., 2015). Fourth, there may be singularities, non-linearities, or structural breaks that prevent invertibility. Our paper does not address these theoretical possibilities, and instead focuses on the empirical question whether the spanning hypothesis is a good description of the data.

Evidence against the spanning hypothesis typically comes from regressions of the form

$$y_{t+h} = \beta_1' x_{1t} + \beta_2' x_{2t} + u_{t+h}, \tag{1}$$

where the dependent variable y_{t+h} is the return or excess return on a long-term bond (or portfolio of bonds), x_{1t} and x_{2t} are vectors containing K_1 and K_2 predictors, respectively, and u_{t+h} is a forecast error. The predictors x_{1t} contain a constant and the information in the yield curve, typically captured by the first three PCs of observed yields, i.e., level, slope, and curvature. The null hypothesis of interest is

$$H_0: \quad \beta_2 = 0,$$

which says that the relevant predictive information is spanned by the information in the yield curve and that x_{2t} has no additional predictive power. A key feature of these regressions is that because the regressors in x_{1t} capture information in the current yield curve, they are necessarily correlated with u_t and hence not strictly exogenous. The predictors are also typically very persistent. This gives rise to a previously unrecognized problem, "standard error bias," that causes tests to reject the null hypothesis much too often, with the problem even more severe when the explanatory variables are trending over the sample. In addition, empirical work typically tries to predict returns over $h = 12$ months, and such use of overlapping returns, and the resulting serial correlation in u_{t+h}, leads to additional econometric problems. In the following subsections we describe these problems in detail.

The spanning hypothesis is of course different from the expectations hypothesis (EH) which posits that expected excess bond returns (i.e., bond risk premia) are constant. A large literature has tested the EH by asking whether any variables help predict excess bond returns. The strongest predictor appears to be the slope of the yield curve, as documented by Fama and Bliss (1987) and Campbell and Shiller (1991). These results are perfectly consistent with the spanning hypothesis. While EH-regressions suffer from small-sample problems similar to those that arise in tests of the spanning regressions, including Stambaugh bias and standard

error bias, Bekaert et al. (1997) and Bekaert and Hodrick (2001) documented that rejections of the EH are robust to accounting for these problems. By contrast, we will show that rejections of the spanning hypothesis are not robust and can arise spuriously.

2.2 The source of standard error bias

Here we explain the intuition for standard error bias in the case when $h = 1$ and u_{t+1} is white noise. According to the Frisch-Waugh Theorem, the OLS estimate of β_2 in (1) can always be viewed as having been obtained in two steps. First we regress x_{2t} on x_{1t} and calculate the residuals $\tilde{x}_{2t} = x_{2t} - \hat{A}_T x_{1t}$ for $\hat{A}_T = \left(\sum_{t=1}^T x_{2t}x_{1t}'\right)\left(\sum_{t=1}^T x_{1t}x_{1t}'\right)^{-1}$. Second we regress y_{t+1} on $\tilde{x}_{2t}$. The coefficient on $\tilde{x}_{2t}$ in this regression will be numerically identical to the coefficient on x_{2t} in the original regression (1).[9] The standard Wald statistic for a test about β_2 can be expressed as

$$W_T = \left(\sum_{t=1}^T u_{t+1}\tilde{x}_{2t}'\right)\left(s^2\sum_{t=1}^T \tilde{x}_{2t}\tilde{x}_{2t}'\right)^{-1}\left(\sum_{t=1}^T \tilde{x}_{2t}u_{t+1}\right) \tag{2}$$

for $s^2 = (T - K_1 - K_2)^{-1}\sum_{t=1}^T (y_{t+1} - b_1'x_{1t} - b_2'x_{2t})^2$ and b_1 and b_2 the OLS estimates from (1). The validity of this test depends on whether W_T is approximately $\chi^2(K_2)$. If x_{1t} and x_{2t} are stationary and ergodic, the estimate $\hat{A}_T$ will converge to the true value $A = E(x_{2t}x_{1t}')\left[E(x_{1t}x_{1t}')\right]^{-1}$. In that case the sampling uncertainty from the first step is asymptotically irrelevant and W would have the same asymptotic distribution as if we replaced $\tilde{x}_{2t}$ with $x_{2t} - Ax_{1t}$, which gives rise to the standard result for stationary regressors that $W_T \xrightarrow{d} \chi^2(K_2)$.

If, however, the regressors are highly persistent, a regression of x_{2t} on x_{1t} behaves like a spurious regression. For example, if x_{1t} and x_{2t} are unit-root processes, the value of $\hat{A}_T$ is not tending to some constant but instead to a random variable $\tilde{A}$ that is different in every sample, even as the sample size T approaches infinity. If x_{1t} was strictly exogenous, this would not affect the asymptotic distribution of W_T. But in tests of the spanning hypothesis x_{1t} is necessarily correlated with u_t, and due to this lack of strict exogeneity $\sum_{t=1}^T \tilde{x}_{2t}u_{t+1}$ has a nonstandard limiting distribution with variance that is larger[10] than that of $\sum_{t=1}^T x_{2t}u_{t+1}$. By contrast, OLS hypothesis tests act as if the variance of $\sum_{t=1}^T \tilde{x}_{2t}u_{t+1}$ is *smaller* than that of $\sum_{t=1}^T x_{2t}u_{t+1}$, since $\sum_{t=1}^T \tilde{x}_{2t}\tilde{x}_{2t}'$ is smaller by construction in every sample than $\sum_{t=1}^T x_{2t}x_{2t}'$. Therefore OLS standard errors are necessarily too small, W_T does not converge to a $\chi^2(K_2)$ distribution, and conventional t- or F-tests about the value of β_2 in (1) will reject more often

[9] We provide a proof of this and other statements in this section in Appendix A.1.

[10] More formally, the difference between the two matrices is a positive definite matrix.

than they should.[11]

2.3 A canonical example

In this section we explore the size of these effects in a canonical example, using first local-to-unity asymptotics and then small-sample simulations based on the model

$$y_{t+1} = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + u_{t+1} \tag{3}$$

where x_{1t} and x_{2t} are scalar AR(1) processes

$$x_{1,t+1} = \mu_1 + \rho_1 x_{1t} + \varepsilon_{1t} \tag{4}$$

$$x_{2,t+1} = \mu_2 + \rho_2 x_{2t} + \varepsilon_{2t} \tag{5}$$

with ε_{it} martingale-difference sequences and $x_{i0} = 0$. Our interest is in what happens when the persistence parameters ρ_i are close to unity. We first focus on the case without drift in these processes ($\mu_1 = \mu_2 = 0$). We assume that innovations to x_{1t} have correlation δ with u_t, whereas x_{2t} is uncorrelated with both x_{1t} and u_t:

$$E \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \\ u_t \end{bmatrix} \begin{bmatrix} \varepsilon_{1s} & \varepsilon_{2s} & u_s \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & 0 & \delta\sigma_1\sigma_u \\ 0 & \sigma_2^2 & 0 \\ \delta\sigma_1\sigma_u & 0 & \sigma_u^2 \end{bmatrix} \quad \text{if } t = s$$
$$= 0 \qquad \text{otherwise.}$$

Thus when $\beta_2 = 0$, the variable x_{2t} has nothing to do with either x_{1s} or y_s for any t or s.

One device for seeing how the results in a finite sample of some particular size T differ from those predicted by conventional first-order asymptotics is to use a local-to-unity specification as in Phillips (1988) and Cavanagh et al. (1995):

$$x_{i,t+1} = (1 + c_i/T)x_{it} + \varepsilon_{i,t+1} \quad i = 1, 2. \tag{6}$$

For example, if our data come from a sample of size $T = 100$ when $\rho_i = 0.99$, the idea is to approximate the small-sample distribution of regression statistics by the asymptotic distribution obtained by taking $c_i = -1$ in (6) and letting $T \to \infty$.[12] The local-to-unity

[11] In Appendix A.1 we go through this argument in more detail, and provide additional proofs. Note also that we have focused on conventional OLS standard errors that assume conditional homoskedasticity, but very similar reasoning applies when White's heteroskedasticity-robust standard errors are used.

[12] It is well known that approximations from such local-to-unity asymptotics are substantially better than

asymptotics turn out to be described by Ornstein-Uhlenbeck processes. For example

$$T^{-2}\textstyle\sum_{t=1}^{T}(x_{it}-\bar{x}_i)^2 \Rightarrow \sigma_i^2 \displaystyle\int_0^1 [J_{c_i}^{\mu}(\lambda)]^2 d\lambda$$

where $\Rightarrow$ denotes weak convergence as $T \to \infty$ and

$$J_{c_i}^{\mu}(\lambda) = J_{c_i}(\lambda) - \int_0^1 J_{c_i}(s)ds \qquad J_{c_i}(\lambda) = c_i \int_0^{\lambda} e^{c_i(\lambda-s)} W_i(s)ds + W_i(\lambda) \qquad i = 1, 2$$

with $W_1(\lambda)$ and $W_2(\lambda)$ denoting independent standard Brownian motion.[13]

We show in Appendix A.2 that under local-to-unity asymptotics the coefficient from a regression of x_{2t} on x_{1t} has the following limiting distribution:

$$A_T = \frac{\sum(x_{1t}-\bar{x}_1)(x_{2t}-\bar{x}_2)}{\sum(x_{1t}-\bar{x}_1)^2} \Rightarrow \frac{\sigma_2 \int_0^1 J_{c_1}^{\mu}(\lambda) J_{c_2}^{\mu}(\lambda) d\lambda}{\sigma_1 \int_0^1 [J_{c_1}^{\mu}(\lambda)]^2 d\lambda} = (\sigma_2/\sigma_1)A, \tag{7}$$

where the last equality defines the random variable A. Under first-order asymptotics the influence of A_T would vanish as the sample size grows. But using local-to-unity asymptotics we see that A_T behaves similarly to the coefficient in a spurious regression and does not converge to zero—the true correlation between x_{1t} and x_{2t} in this setting—but to a random variable that differs across samples. The implication is that the t-statistic for b_2 can have a small-sample distribution that is very poorly approximated using first-order asymptotics. Appendix A.2 demonstrates that this t-statistic has a local-to-unity asymptotic distribution under the null hypothesis that is given by

$$\frac{b_2 - \beta_2}{\{s^2/\sum \tilde{x}_{2t}^2\}^{1/2}} \Rightarrow \delta Z_1 + \sqrt{1-\delta^2} Z_0 \tag{8}$$

$$Z_1 = \frac{\int_0^1 K_{c_1,c_2}(\lambda) dW_1(\lambda)}{\left\{\int_0^1 [K_{c_1,c_2}(\lambda)]^2 d\lambda\right\}^{1/2}} \qquad Z_0 = \frac{\int_0^1 K_{c_1,c_2}(\lambda) dW_0(\lambda)}{\left\{\int_0^1 [K_{c_1,c_2}(\lambda)]^2 d\lambda\right\}^{1/2}} \qquad K_{c_1,c_2}(\lambda) = J_{c_2}^{\mu}(\lambda) - A J_{c_1}^{\mu}(\lambda)$$

for $s^2 = (T-3)^{-1}\sum(y_{t+1} - b_0 - b_1 x_{1t} - b_2 x_{2t})^2$ and $W_i(\lambda)$ independent standard Brownian processes for $i = 0, 1, 2$. Conditional on the realizations of $W_1(.)$ and $W_2(.)$, the term Z_0 will be recognized as a standard Normal variable, and therefore Z_0 has an unconditional $N(0,1)$ distribution as well.[14] In other words, if x_{1t} is strictly exogenous ($\delta = 0$) then the OLS t-test

those based on conventional first-order asymptotics which take $T \to \infty$ and treat $\rho_i = 0.99$ as a constant; see for example Chan (1988) and Nabeya and Sørensen (1994).

[13] When $c_i = 0$, (6) becomes a random walk and the local-to-unity asymptotics simplify to the standard unit-root asymptotics involving functionals of Brownian motion as a special case: $J_0(\lambda) = W(\lambda)$.

[14] The intuition is that for $v_{0,t+1} \sim$ i.i.d. $N(0,1)$ and $K = \{K_t\}_{t=1}^{T}$ any sequence of random variables

of $\beta_2 = 0$ will be valid in small samples even with highly persistent regressors. By contrast, if $\delta \neq 0$ the random variable Z_1 comes into play, which has a nonstandard distribution because the term $dW_1(\lambda)$ in the numerator is not independent of the denominator. In particular, Appendix A.2 establishes that $\text{Var}(Z_1) > 1$. Moreover Z_1 and Z_0 are uncorrelated with each other.[15] Therefore the t-statistic in (8) has a non-standard distribution with variance $\delta^2\text{Var}(Z_1) + (1-\delta^2)1 > 1$ which is monotonically increasing in $|\delta|$. This shows that whenever x_{1t} is correlated with u_t ($\delta \neq 0$) and x_{1t} and x_{2t} are highly persistent, in small samples the t-test of $\beta_2 = 0$ will reject too often when H_0 is true.[16]

We can quantify the magnitude of these effects in a simulation study. We generate values for x_{1t} and x_{2t} by drawing ε_{1t} and ε_{2t} as i.i.d. Gaussian random variables with $\sigma_1 = \sigma_2 = 1$, using $\mu_1 = \mu_2 = 0$ and different values of $\rho_1 = \rho_2 = \rho$, starting from $x_{10} = x_{20} = 0$. We generate $y_t = u_t = \delta\varepsilon_{1t} + \sqrt{1-\delta^2}v_t$ where v_t is a standard normal random variable.[17] Hence, in our data-generating process (DGP) we have $\beta_0 = \beta_1 = \beta_2 = 0$, $\sigma_u = 1$, and $Corr(u_t, \varepsilon_{1t}) = \delta$. We simulate 1,000,000 samples, estimate regression (3) in each sample, and study the small-sample behavior of the t-statistic for the test of $H_0 : \beta_2 = 0$, using critical values from the Student-t distribution with 97 degrees of freedom. In addition, we also draw from the local-to-unity asymptotic distribution of the t-statistic given in equation (8) using well-known Monte Carlo methods.[18]

The first panel of Table 1 shows the results of this exercise for different values of ρ and δ. If the regressors are either strictly exogneous ($\delta = 0$) or not serially correlated ($\rho = 0$), the true size of the t-test of $\beta_2 = 0$ is equal to the nominal size of five percent. If, however, both $\rho \neq 0$ and $\delta \neq 0$, the true size exceeds the nominal size, and this size distortion increases in ρ

---

that is independent of v_0, $\sum_{t=1}^{T} K_t v_{0,t+1}$ has a distribution conditional on K that is $N(0, \sum_{t=1}^{T} K_t^2)$ and $\sum_{t=1}^{T} K_t v_{0,t+1} / \sqrt{\sum_{t=1}^{T} K_t^2} \sim N(0,1)$. Multiplying by the density of K and integrating over K gives the identical unconditional distribution, namely $N(0,1)$. For a more formal discussion in the current setting, see Hamilton (1994, pp. 602-607).

[15] The easiest way to see this is to note that conditional on $W_1(.)$ and $W_2(.)$ the product has expectation zero, so the unconditional expected product is zero as well.

[16] Expression (8) can be viewed as a straightforward generalization of result (2.1) in Cavanagh et al. (1995) and expression (11) in Campbell and Yogo (2006). In their case the explanatory variable is $x_{1,t-1} - \bar{x}_1$ which behaves asymptotically like $J_{c_1}^{\mu}(\lambda)$. The component of u_t that is correlated with ε_{1t} leads to a contribution to the t-statistic given by the expression that Cavanagh et al. (1995) refer to as τ_{1c}, which is labeled as τ_c/κ_c by Campbell and Yogo (2006). This variable is a local-to-unity version of the Dickey-Fuller distribution with well-known negative bias. By contrast, in our case the explanatory variable is $\tilde{x}_{2.t-1} = x_{2,t-1} - A_T x_{1,t-1}$ which behaves asymptotically like $K_{c_1,c_2}(\lambda)$. Here the component of u_t that is correlated with ε_{1t} leads to a contribution to the t-statistic given by Z_1 in our expression (8). Unlike the Dickey-Fuller distribution, Z_1 has mean zero, so that there is no bias in b_2.

[17] We can focus on $0 \leq \delta \leq 1$, since only $|\delta|$ matters for the distribution of the t-statistic.

[18] We simulate samples of size $\tilde{T}$ from near-integrated processes with $c_1 = c_2 = T(\rho - 1)$ and approximate the integrals in (8) using Rieman sums—see, for example, Chan (1988), Stock (1991), and Stock (1994). We use $\tilde{T} = 1000$, since even moderate sample sizes generally yield accurate approximations to the limiting distribution (Stock, 1991, uses $\tilde{T} = 500$).

and δ. In the presence of high persistence the true size of this t-test can be quite substantially above the nominal size: For $\rho = 0.99$ and $\delta = 1$, the true size is around 15 percent, meaning that we would reject the true null hypothesis more than three times as often as we should. The size calculations are, not surprisingly, very similar for the small-sample simulations and the local-to-unity asymptotic approximations.

Figure 1 plots the size of the t-test for the case with $\delta = 1$ for sample sizes from $T = 50$ to 1000, based on the local-to-unity approximation. When $\rho < 1$, the size distortion decreases with the sample size. For example for $\rho = 0.99$ the size decreases from 15 percent to about 9 percent. In contrast, when $\rho = 1$ the size distortions are not affected by the sample size, as indeed in this case the non-Normal distribution corresponding to (8) with $c_i = 0$ governs the distribution for arbitrarily large T.

The reason for the size distortions when testing $\beta_2 = 0$ is not coefficient bias. The top panel of Table 1 shows that b_1 is downward biased but b_2 is unbiased. However, the conventional OLS standard errors underestimate the true sampling variability of the OLS estimates: they can average up to 30% below the standard deviation of the coefficient estimates across simulations. This standard error bias is the reason why the t-test rejects too often.

2.4 The role of trends

Up to now we have been considering the case when the true values of the constant terms μ_i in equations (4)-(5) are zero. As seen in the second and third panels of Table 1, the size distortions on tests about β_2 can nearly double when $\mu_i \neq 0$, and the bias in the estimate of β_1 increases as well.

We can understand what is going on most easily by considering the case when the roots ρ_i are exactly unity.[19] In that case, if μ_1 is zero and μ_2 is not, x_{2t} will exhibit a deterministic time trend and this ends up stochastically dominating the random walk component of x_{2t}. The regression (3) would then be asymptotically equivalent to a regression of y_{t+1} on $(1, x_{1t}, \mu_2 t)'$. When $\delta = 1$ the asymptotic distribution of a t-test of a true null hypothesis about β_1 in regression (3) would be identical to that if we were to perform a Dickey-Fuller test of the true null hypothesis $\eta = 0$ in the regression

$$\Delta x_{1,t+1} = \mu_1 + \eta x_{1t} + \xi t + \varepsilon_{1,t+1}, \tag{9}$$

which is the well-known Dickey-Fuller Case 4 distribution described in Hamilton (1994, eq. [17.4.55]). We know that the coefficient bias and size distortions are bigger when a time trend

[19] The following results are proved formally in Appendix A.3.

is included in regression (9) compared to the case when it is not.[20] For the same reason we would find that the Stambaugh bias of b_1 in regression (3) becomes worse when a variable x_{2t} with a deterministic trend is added to the regression. The standard error bias for b_2 is also exacerbated when the true μ_2 is nonzero.

In the case when ρ_2 is close to but strictly less than unity, this problem would vanish asymptotically but is still a factor in small samples. An apparent trend shows up in a finite sample because when a stationary variable with mean $\mu_2/(1-\rho_2)$ is started from $x_{20} = 0$, it will tend to rise toward its unconditional mean. As ρ_2 approaches unity, this trend within a finite sample becomes arbitrarily close to that seen in a true random walk with drift μ_2. While in the applications in this paper the trend in explanatory variables like inflation is typically downward instead of upward, the issue is the same, because the distribution of b_1 is identical whether x_{t2} begins at $x_{20} = 2\mu_2/(1-\rho_2)$ and then drifts down to its mean or whether it begins at $x_{20} = 0$ and drifts up. Note that the values we have used for simulation in Table 1 are representative of those that may be encountered in practice.[21]

When both x_{1t} and x_{2t} have trends (see panel 3 of Table 1) we have the same issues just discussed but with a reinterpretation of the variables. Consider for example the case when both trends are the same ($\mu_1 = \mu_2$). Note that a regression of y_{t+1} on $(1, x_{1t}, x_{2t})'$ has the identical fitted values as a regression of y_{t+1} on $(1, x_{1t} - x_{2t}, x_{2t})'$, which again is asymptotically equivalent to a regression in which the second variable is a driftless unit-root process correlated with the lagged residual and the third variable is dominated by a deterministic time trend. Now the Stambaugh bias will show up in the coefficient on $x_{1t} - x_{2t}$. Translating back in terms of the original regression of y_{t+1} on $(1, x_{1t}, x_{2t})'$ we would now find Stambaugh biases in both b_1 and b_2 that are mirror images of each other. Note the implications of this example. When μ_1 and μ_2 are both nonzero, if we were to regress y_{t+1} on x_{1t} alone, there would be no Stambaugh bias and no problem with t-tests about β_1, because x_{1t} is dominated by the time trend. The same is true if we were to regress y_{t+1} on x_{2t} alone. But when both x_{1t} and x_{2t} are included in the regression, spurious conclusions about both coefficients would emerge.

The practical relevance of these results is that when the proposed additional predictors in x_{2t} are trending, this can substantially magnify the small-sample problems and lead to more poorly sized tests and spurious rejections of the spanning hypothesis.

[20] See Case 2 versus Case 4 in Hamilton (1994, Tables B.5 and B.6).

[21] For example, an AR(1) process fit to the trend inflation variable used by Cieslak and Povala (2015) over the sample 1985-2013 has $\rho_2 = 0.99$ and $\mu_2/\sigma_2 = 1.5$, an even stronger drift relative to innovation than the value $\mu_2/\sigma_2 = 1.0$ used in Table 1. And their variable has a value in 1985:1 that is 5 times the size of $\mu_2/(1-\rho_2)$, implying a downward drift over 1985-2013 that is 4 times as fast as in the Table 1 simulation.

2.5 Overlapping returns

A separate econometric problem arises in predictive regressions for bond returns with holding periods that are longer than the sampling interval, i.e., $h > 1$. Most studies in this literature, and all those that we revisit in this paper, focus on predictive regressions for annual excess bond returns in monthly data, that is regression (1) with $h = 12$ and

$$y_{t+h} = p_{n-h,t+h} - p_{nt} - hi_{ht}, \tag{10}$$

for p_{nt} the log of the price of a pure discount n-period bond purchased at date t and $i_{nt} = -p_{nt}/n$ the corresponding zero-coupon yield. In that case, $E(u_t u_{t-v}) \neq 0$ for $v = 0, \ldots, h-1$, as the overlapping observations induce a MA$(h-1)$ structure for the error terms. This raises additional problems in the presence of persistent regressors that can be seen even using conventional first-order asymptotics, as we briefly note in this section.

If x_{1t} and x_{2t} are uncorrelated and the true value of $\beta_2 = 0$, we show in Appendix A.4 that under conventional first-order asymptotics

$$\sqrt{T} b_2 \xrightarrow{d} N(0, Q^{-1} S Q^{-1}), \tag{11}$$

$$Q = E(x_{2t} x_{2t}'), \quad S = \sum\nolimits_{v=-\infty}^{\infty} E(u_{t+h} u_{t+h-v} x_{2t} x_{2,t-v}'). \tag{12}$$

Note that the product $u_{t+h} x_{2t}$ will be serially correlated when x_{2t} is persistent, since $E(u_{t+h} u_{t+h-v} x_{2t} x_{2,t-v}') = E(u_t u_{t-v}) E(x_{2t} x_{2,t-v}') \neq 0$. Overlapping observations, in combination with persistent regressors, substantially increase the sampling variability of the OLS estimate b_2, because the long-run covariance matrix S will exceed the value $S_0 = E(u_{t+h}^2 x_{2t} x_{2t}')$ that would be appropriate for serially uncorrelated residuals.

The standard approach is to use heteroskedasticity- and autocorrelation-consistent (HAC) standard errors to try to correct for this, for example, the estimators proposed by Newey and West (1987) or Andrews (1991). However, long-run variance estimation is notoriously difficult, particularly in small samples, and different HAC estimators of S can lead to substantially different empirical conclusions (Müller, 2014). That Newey-West standard errors are unreliable for inference with overlapping returns was demonstrated convincingly by Ang and Bekaert (2007). Here we emphasize that the higher the persistence of the predictors, the less reliable is HAC inference, since the effective sample size becomes very small. The reverse-regression approach of Hodrick (1992) and Wei and Wright (2013) can alleviate but not overcome the problem arising from overlapping returns, as we will show in Section 5.

There is another consequence of basing inference on overlapping observations that appears not to be widely recognized: it substantially reduces the reliability of R^2 as a measure of

goodness of fit. Let R_1^2 denote the coefficient of determination in a regression that includes only x_{1t}, compared to R_2^2 for the regression that includes both x_{1t} and x_{2t}. We show in Appendix A.4 that again for the case when x_{1t} and x_{2t} are uncorrelated and $\beta_2 = 0$

$$T(R_2^2 - R_1^2) \xrightarrow{d} r'Q^{-1}r/\gamma, \quad \gamma = E[y_t - E(y_t)]^2, \quad r \sim N(0, S). \tag{13}$$

The difference $R_2^2 - R_1^2$ converges in probability to zero, but in a given finite sample it is positive by construction. If $x_{2t}u_{t+h}$ is positively serially correlated, then S exceeds S_0 by a positive-definite matrix, and r exhibits more variability across samples. This means $R_2^2 - R_1^2$, being a quadratic form in a vector with a higher variance, would have both a higher expected value as well as a higher variance when $x_{2t}u_{t+h}$ is serially correlated compared to situations when it is not. This serial correlation in $x_{2t}u_{t+h}$ would contribute to larger values for $R_2^2 - R_1^2$ on average as well as to increased variability in $R_2^2 - R_1^2$ across samples. In other words, including x_{2t} could substantially increase the R^2 even if H_0 is true. We will use bootstrap approximations to the small-sample distribution of $R_2^2 - R_1^2$, and demonstrate that the dramatic values sometimes reported in the literature are often entirely plausible under the spanning hypothesis.

2.6 A bootstrap design to test the spanning hypothesis

Obviously the main question is whether the above considerations make a material difference for tests of the spanning hypothesis. We propose a parametric bootstrap that generates data under the spanning hypothesis to assess how serious these econometric problems are in practice.[22] With this bootstrap approach we can calculate the size of conventional tests to assess their robustness. In addition, we can use it to test the spanning hypothesis with better size and power than for conventional tests.[23]

Our bootstrap design is as follows: First, we calculate the first three PCs of observed yields which we denote

$$x_{1t} = (PC1_t, PC2_t, PC3_t)',$$

along with the weighting vector $\hat{w}_n$ for the bond yield with maturity n:

$$i_{nt} = \hat{w}_n' x_{1t} + \hat{v}_{nt}.$$

[22] An alternative approach would be a nonparametric bootstrap under the null hypothesis, using for example a moving-block bootstrap to re-sample x_{1t} and x_{2t}. However, Berkowitz and Kilian (2000) found that parametric bootstrap methods such as ours typically perform better than nonparametric methods.

[23] Cochrane and Piazzesi (2005) and Ludvigson and Ng (2009, 2010) also used the bootstrap to test $\beta_2 = 0$. They did so with bootstrap confidence intervals generated under the alternative hypothesis. But it is well known that bootstrapping under the null hypothesis generally leads to better numerical accuracy and more powerful tests (Hall and Wilson, 1991; Horowitz, 2001), and of course this is the only way to obtain bootstrap estimates of the size of conventional tests.

That is, $x_{1t} = \hat{W} i_t$, where $i_t = (i_{n_1 t}, \ldots, i_{n_J t})'$ is a J-vector with observed yields at t, and $\hat{W} = (\hat{w}_{n_1}, \ldots, \hat{w}_{n_J})'$ is the $3 \times J$ matrix with rows equal to the first three eigenvectors of the variance matrix of i_t. We use normalized eigenvectors so that $\hat{W}\hat{W}' = I_3$. Fitted yields are obtained as $\hat{\imath}_t = \hat{W}' x_{1t}$. Three factors generally fit the cross section of yields very well, with fitting errors $\hat{v}_{nt}$ (pooled across maturities) that have a standard deviation of only a few basis points.[24] Then we estimate by OLS a VAR(1) for x_{1t}:

$$x_{1t} = \hat{\phi}_0 + \hat{\phi}_1 x_{1,t-1} + e_{1t} \qquad t = 1, \ldots, T. \tag{14}$$

This time-series specification for x_{1t} completes our simple factor model for the yield curve. Though this model does not impose absence of arbitrage, it captures both the dynamic evolution and the cross-sectional dependence of yields. A no-arbitrage model is a special case of this structure with additional restrictions on $\hat{W}$, but these restrictions typically do not improve forecasts of yields; see for example Duffee (2011a) and Hamilton and Wu (2014). Next we generate 5000 artificial yield data samples from this model, each with length T equal to the original sample length. We first iterate on

$$x^*_{1\tau} = \hat{\phi}_0 + \hat{\phi}_1 x^*_{1,\tau-1} + e^*_{1\tau}$$

where $e^*_{1\tau}$ denotes bootstrap residuals. We start every bootstrap sample at $x^*_{10} = x_{10}$, the starting value for the observed sample, to allow for a possible contribution of trends resulting from initial conditions as discussed in Section 2.4. Then we obtain the bootstrap yields using

$$i^*_{n\tau} = \hat{w}'_n x^*_{1\tau} + v^*_{n\tau} \tag{15}$$

for $v^*_{n\tau} \overset{iid}{\sim} N(0, \sigma_v^2)$. The standard deviation of the measurement errors, σ_v, is set to the sample standard deviation of the fitting errors $\hat{v}_{nt}$.[25] We thus have generated an artificial sample of yields $i^*_{n\tau}$ which by construction only the three factors in $x^*_{1\tau}$ have any power to predict, but whose covariance and dynamics are similar to those of the observed data i_{nt}.

We likewise fit a VAR(1) to the observed data for the proposed predictors x_{2t},

$$x_{2t} = \hat{\alpha}_0 + \hat{\alpha}_1 x_{2,t-1} + e_{2t}, \tag{16}$$

from which we then bootstrap 5000 artificial samples $x^*_{2\tau}$ in a similar fashion as for $x^*_{1\tau}$. The

[24] For example, in the data of Joslin et al. (2014) this standard deviation is 6.5 basis points.

[25] Some evidence in the literature suggests that yield fitting errors are serially correlated (Adrian et al., 2013; Hamilton and Wu, 2014). We have also investigated a setting with serial correlation in $v^*_{n\tau}$ and found that this does not change any of our findings.

bootstrap residuals $(e'^{*}_{1\tau}, e'^{*}_{2\tau})$ are drawn from the joint empirical distribution of (e'_{1t}, e'_{2t}).

Using the bootstrapped samples of predictors and yields, we can then investigate the properties of any proposed test statistic involving $y^*_{\tau+h}$, $x^*_{1\tau}$, and $x^*_{2\tau}$ in a sample in which the serial correlation of these variables is similar to the actual data, but in which by construction the null hypothesis is true that $x^*_{2\tau}$ has no incremental predictive power.[26] Consider for example a t-test for significance of a parameter in β_2. Denote the t-statistic in the data by t and the corresponding t-statistic in bootstrap sample i as t^*_i. To obtain a bootstrap estimate of the size of this test we calculate the fraction of samples in which $|t^*_i|$ exceeds the usual asymptotic critical value. And to use the bootstrap to carry out the hypothesis test, we calculate the bootstrap p-value as the fraction of samples in which $|t^*_i| > |t|$, and reject the null if this is less than, say, five percent. Equivalently, we can calculate the bootstrap critical value as the 95th percentile of $|t^*_i|$ and reject the null if $|t|$ exceeds it.

Note that this bootstrap procedure does not generate a test with an exact size of 5%. First, under local-to-unity asymptotics the bootstrap is not a consistent test because the test statistics are not asymptotically pivotal—their distribution depends on the nuisance parameters c_1 and c_2, which cannot be consistently estimated.[27] Second, least squares typically underestimates the autocorrelation of highly persistent processes due to small-sample bias (Kendall, 1954; Pope, 1990), so that the VAR underlying our bootstrap would typically be less persistent than the true DGP. We can address the second issue by using bias-corrected VAR parameter estimates for generating bootstrap samples. We will use the bias correction proposed by Kilian (1998) and refer to this as the "bias-corrected bootstrap."[28] We have found that even the bias-corrected bootstrap tends to be slightly oversized. This means that if our bootstrap test fails to reject the spanning hypothesis, the reason is not that the test is too conservative, but that there simply is not sufficient evidence for rejecting the null.

We can use the Monte Carlo simulations in Section 2.3 to calculate the size of our bootstrap test. In each sample i simulated from a known parametric model, we can: (i) calculate the t-statistic (denoted $\tilde{t}_i$) for testing the null hypothesis that $\beta_2 = 0$; (ii) estimate the autoregressive models for the predictors by using OLS on that sample; (iii) generate a *single* bootstrap sample using these estimated autoregressive coefficients; (iv) estimate the predictive regression on

[26] For example, if y_{t+h} is an h-period excess return as in equation (10) then in our bootstrap

$$\begin{aligned} y^*_{\tau+h} &= ni^*_{n\tau} - hi^*_{h\tau} - (n-h)i^*_{n-h,\tau+h} \\ &= n(\hat{w}'_n x^*_{1\tau} + v^*_{n\tau}) - h(\hat{w}'_h x^*_{1\tau} + v^*_{h\tau}) - (n-h)(\hat{w}'_{n-h} x^*_{1,\tau+h} + v^*_{n-h,\tau+h}) \end{aligned}$$

which replicates the predictable component and the MA$(h-1)$ serial correlation structure of the excess returns that is both seen in the data and predicted under the spanning hypothesis.

[27] This result goes back to Basawa et al. (1991). See also Horowitz (2001) and the references therein.

[28] We have found in Monte Carlo experiments that the size of the bias-corrected bootstrap is closer to five percent than for the simple bootstrap.

the bootstrap sample;[29] and (v) calculate the t-statistic in this regression, denoted t_i^*. We generate many samples from the maintained model, repeating steps (i)-(v), and then calculate the value c such that $|t_i^*| > c$ in 5% of the samples. Our bootstrap procedure amounts to the recommendation of rejecting H_0 if $|\tilde{t}_i| > c$, and we can calculate from the above simulation the fraction of samples in which this occurs. This number tells us the true size if we were to apply our bootstrap procedure to the chosen parametric model. This number is reported in the last column of Table 1. We find in these settings that our bootstrap has a size above but fairly close to five percent.

We will repeat the above procedure to estimate the size of our bootstrap test in each of our empirical applications, taking a model whose true coefficients are those of the VAR estimated in the sample as if it were the known parametric model, and estimating VAR's from data generated using those coefficients. To foreshadow those results, we will find that the size is typically quite close to or slightly above five percent, and that our bootstrap procedure has excellent power. The implication is that if our bootstrap procedure fails to reject the spanning hypothesis, we should conclude that the evidence against the spanning hypothesis in the original data is not persuasive.

2.7 New data: subsample stability and out-of-sample forecasting

We also reassess reported claims of violations of the spanning hypothesis by confronting them with new data released after publication of the original studies. To circumvent econometric problems of predictability regressions a common practice is to perform pseudo out-of-sample (OOS) analysis, splitting the sample into an initial estimation and an OOS period. We are skeptical of this approach because the researcher has access to the full sample when formulating the model, and the sample-split is arbitrary. However, for each of the studies that we revisit a significant amount of new data have come in since the original research. This gives us an opportunity both to reestimate the models over a sample period that includes new data, and further to evaluate the true out-of-sample forecasting performance of each proposed model.

[29] In this simple Monte Carlo setting, we bootstrap the dependent variable as $y_\tau^* = \hat{\phi}_1 x_{1,\tau-1}^* + u_\tau^*$ where u_τ^* is resampled from the residuals in a regression of y_t on $x_{1,t-1}$, and is jointly drawn with $\varepsilon_{1\tau}^*$ and $\varepsilon_{2\tau}^*$ to maintain the same correlation as in the data. By contrast, in our empirical analysis the bootstrapped dependent variable is calculated from the bootstrapped bond yields, obtained using (15), and the definition of y_{t+h} (for example, as an annual excess return).

3 Economic growth and inflation

In this section we examine the evidence reported by Joslin et al. (2014) (henceforth JPS) that macro variables may help predict bond returns. We will follow JPS and focus on predictive regressions as in equation (1) where y_{t+h} is an excess bond return for a one-year holding period ($h = 12$), x_{1t} is a vector consisting of a constant and the first three PCs of yields, and x_{2t} consists of a measure of economic growth (the three-month moving average of the Chicago Fed National Activity Index, GRO) and of inflation (one-year CPI inflation expectations from the Blue Chip Financial Forecasts, INF). While JPS also presented model-based evidence in favor of unspanned macro risks, those results stem from the substantial in-sample predictive power of x_{2t} in the excess return regressions. The sample contains monthly observations over the period 1985:1-2008:12.[30]

3.1 Predictive power according to $\bar{R}^2$

JPS found that for the ten-year bond, the (adjusted) $\bar{R}^2$ of regression (1) increased from 0.20 to 0.37 when x_{2t} is included. For the two-year bond, the change is even more striking, with $\bar{R}^2$ increasing from 0.14 to 0.48. JPS interpreted this as strong evidence that macroeconomic variables have predictive power for excess bond returns beyond the information in the yield curve, and concluded that "macroeconomic risks are unspanned by bond yields" (p. 1203). We report the $\bar{R}^2$ for an average excess return on 2- to 10-year bonds in the first row of Table 2, where the first three entries are based on the same data set that was used by JPS.[31] The entry $\bar{R}^2_1$ gives the $\bar{R}^2$ for the regression with only x_{1t} as predictors, and $\bar{R}^2_2$ corresponds to the case when x_{2t} is added to the regression. For this specification, $\bar{R}^2$ also increases quite substantially, by 19 percentage points.

However, there are some warning flags for these predictive regressions. First, the predictors are very persistent; the first-order sample autocorrelations of $PC1$ and $PC2$ are 0.98 and 0.97, respectively, while that of INF is 0.99. Second, the sample is relatively small, with 276 observations. Third, the dependent variable is an annual overlapping return, i.e., $h = 12$. The arguments in Section 2.5 therefore suggest that even large increases in $\bar{R}^2$ may be plausible

[30] We recreated the data set using unsmoothed Fama-Bliss yields from Anh Le (Le and Singleton, 2013) and data from the Chicago Fed and Blue Chip to reconstruct GRO and INF. Note that the last observation corresponds to excess returns over the holding period from 2007:12 to 2008:12.

[31] In Table 2 we have attempted to summarize results for R^2 or $\bar{R}^2$ across different studies on a comparable basis that is as close as possible to that in the original study. In the case of JPS, they reported results for only the 2-year and 10-year bonds and not an average. In Table C.1 in Appendix C we present analogous results for each individual bond from two through ten years maturity. The increase in $\bar{R}^2$ when adding macro variables is particularly pronounced for short-term bonds, but most of our conclusions apply to these short maturities as well.

under the null hypothesis.

The second row of Table 2 reports the mean $\bar{R}^2$ across 5000 replications of the bootstrap described in Section 2.6, that is, the average value we would expect to see for these statistics in a sample of the size used by JPS in which x_{2t} in fact has no true ability to predict y_{t+h} but whose serial correlation properties are similar to those of the observed data. The third row gives 95% bootstrap intervals, that is, the 2.5th and 97.5th percentiles of the bootstrap distributions which impose the null hypothesis. The variability of the $\bar{R}^2$ is very high. Values for $\bar{R}^2_2$ as high as 60% would not be uncommon, as indicated by the bootstrap intervals. Most notably, adding the regressors x_{2t} often substantially increases the $\bar{R}^2$—even increases of 20 percentage points are not uncommon—although x_{2t} has no predictive power in population by construction. According to the bootstrap small-sample distribution of $\bar{R}^2$, the increase in the data of 19 percentage points is not inconsistent with the spanning hypothesis.

Since the persistence of x_{2t} is high, it may be important to adjust for small-sample bias in the VAR estimates, so we also carried out the bias-corrected (BC) bootstrap. The expected values and 95% bootstrap intervals are reported in the bottom two rows of the top panel in Table 2. As expected, more serial correlation in the generated data (due to the bias correction) increases the mean and the variability of the $\bar{R}^2$ and of their difference. Hence $\bar{R}^2_2 - \bar{R}^2_1$ is even more comfortably within the bootstrap interval.

3.2 Testing the spanning hypothesis

While JPS only reported $\bar{R}^2$ for their excess return regression, one is naturally interested in formal tests of the spanning hypothesis. We report coefficient estimates and test statistics in Table 3. The common approach to address the serial correlation in the residuals due to overlapping observations is to use the standard errors and test statistics proposed by Newey and West (1987), and in regressions for annual returns with monthly data researchers typically use 18 lags (see among many others Cochrane and Piazzesi, 2005; Ludvigson and Ng, 2009). In the second row of Table 3 we report the resulting t-statistic for each coefficient along with the Wald test of the hypothesis $\beta_2 = 0$. The third row reports the p-values for these statistics if they were interpreted using the conventional asymptotic approximation. According to this popular test, GRO and INF appear strongly significant, both individually and jointly. In particular, the Wald test gives a p-value below 0.1%.

However, the small-sample problems described in Section 2 likely distort these test results. The canonical correlation between innovations in one-month excess returns and innovations in the three yield PCs (the generalization of the parameter δ in Section 2.3) is 0.99. This correlation is always high in tests of the spanning hypothesis, because the yield PCs in x_{1t}

explain current yields very well, and so innovations to x_{1t} are highly correlated with returns realized at t. Furthermore, as noted above, the autocorrelations of the predictors are high and the sample size is relatively small. Our theory predicts that standard error bias will be severe in this application. In addition, the well-known small-sample problems of Newey-West standard errors are likely to be particularly pronounced in this setting.

We therefore employ our bootstrap to carry out tests of the spanning hypothesis that account for these small-sample issues. Again, we use both simple (OLS) and BC bootstrap. For each, we report five-percent critical values for the t- and Wald statistics, calculated as the 95th percentiles of the bootstrap distribution, as well as bootstrap p-values, i.e., the frequency of bootstrap replications in which the bootstrapped test statistics are at least as large as in the data. Using either the simple or BC bootstrap, the coefficient on GRO is insignificant even at the 10% level, and the coefficient on INF is marginally significant at the 5% level. The bootstrap p-value for the Wald test of the spanning hypothesis is slightly below 5% for the simple bootstrap and slightly above 5% for the BC bootstrap. These tests give much weaker evidence against the spanning hypothesis than one would have thought based on conventional asymptotic interpretation of the test statistics.

Using the bootstrap we can calculate the true size of the conventional HAC and the bootstrap tests, which both have a nominal size of five percent. These are reported in the *Size* section of the top panel of Table 3. For the conventional HAC tests, this is calculated as the frequency of bootstrap replications in which the test statistics exceed the usual asymptotic critical values. The results reveal that the true size of these conventional tests is 19-36% instead of the presumed five percent. These substantial size distortions are also reflected in the bootstrap critical values, which far exceed the conventional ones.

We can also use our bootstrap to evaluate the power of our proposed tests. To do so, we simply add $\hat{\beta}_2 x^*_{2\tau}$ to the value generated by our bootstrap for $y^*_{\tau+h}$, where $\hat{\beta}_2$ is the coefficient on x_{2t} in the original data sample. We now have a generated sample in which x_{2t} in fact does predict y_{t+h}, and with a magnitude that is exactly that claimed in the original study. We repeat this to obtain 5000 such samples and in each sample calculate all our tests. We find that the bootstrap Wald test rejects the (false) spanning hypothesis in 89% of the samples. In other words, these tests should reject the spanning hypothesis in the data if it were indeed false, which suggests that the reason that they do not reject is not a lack of power, but the fact that empirical spanning is a reasonable description of the observed sample.

In addition we also tested alternative versions of the spanning hypothesis where four or five PCs of yields capture the information in the yield curve. The results, reported in Appendix B, show that our conclusions are unchanged when we allow for a more general spanning hypothesis.

3.3 New data

What happens when we augment the sample with the eight years of new data that have arrived since the original analysis by JPS? The last three columns of the top panel of Table 2 show that the in-sample improvement in $\bar{R}^2$ when x_{2t} is included in the regression is substantially smaller over the 1985-2016 data set than was found on the original JPS data set, and the improvement is far from statistically significant.[32] And as seen in the second panel of Table 3, the values of the HAC test statistics are substantially smaller on the longer data set than in the original data, and the t- and Wald statistics are no longer statistically significant even if interpreted in the usual way.

Row 1 of Table 4 reports the pure OOS forecast comparison for y_{t+h} the average 12-month excess return across 2- to 10-year bonds. We used a recursive scheme where we re-estimate the predictive regressions by extending the estimation window each month of the newly available data. Whereas in the original JPS in-sample regression, the addition of x_{2t} improved the mean squared prediction error by 24%, the addition of x_{2t} leads to a deterioration in the OOS prediction error by 116%. Moreover, this deterioration is strongly statistically significant according to the Diebold and Mariano (1995) (DM) test.[33]

Adding new observations to the JPS data set substantially weakens the evidence against the spanning hypothesis. But if the null hypothesis were truly false, we would expect to find the evidence against it become stronger, not weaker, when we use a bigger data set. We conclude on the basis of the bootstrap and the evidence in newly available data that the JPS evidence on unspanned macro risks is far from convincing.

4 Factors of large macro data sets

Ludvigson and Ng (2009, 2010) found that factors extracted from a large macroeconomic data set are helpful in predicting excess bond returns, above and beyond the information contained in the yield curve. Here we revisit this evidence, focusing on the results in Ludvigson and Ng (2010) (henceforth LN). They started with a panel data set of 131 macro variables observed over 1964:1-2007:12 and extracted eight macro factors using the method of principal components. These factors, which we will denote by $F1$ through $F8$, were then related to future one-year excess returns on two- through five-year Treasury bonds. They also included

[32] This also turns out to be the case for every individual bond maturity; see Table C.1 in Appendix C.

[33] In related work, Giacoletti et al. (2016) evaluated the real-time OOS forecasting performance of a model similar to that used in JPS. They found that including macro variables only helps for predicting very short-term yields and only over a specific subsample, but that overall "'macro rules' add little to the forecast accuracy of the basic yields-only rule" (p. 29). While this supports the spanning hypothesis, they find some incremental predictive power when including survey forecast disagreement.

the return-forecasting factor that was proposed by Cochrane and Piazzesi (2005), denoted CP, which is the linear combination of forward rates that best predicts the average excess return across maturities. Based on comparisons of $\bar{R}^2$ of regressions with and without macro factors, as well as inference using Newey-West standard errors, LN concluded that macro factors help predict excess returns, even when controlling for information in the yield curve using the CP factor.

We estimate regression (1) where now y_{t+h} is the average one-year excess bond return for maturities of two through five years, x_{1t} contains a constant and three yield PCs, and x_{2t} contains eight macro PCs. The specification is very similar to that of LN, with two differences: First, we capture the information in the yield curve using the first three PCs of yields, while LN use the CP factor. Second, we do not carry out LN's preliminary specification search—they considered many different combinations of the factors along with squared and cubic terms—in order to focus squarely on hypothesis testing for a given regression specification.[34]

Table 2 shows that in LN's data set the $\bar{R}^2$ increases by 10 percentage points when the macro factors are included, consistent with LN's findings. The first three rows of Table 5 show the coefficient estimates, HAC t- and Wald statistics (using Newey-West standard errors with 18 lags as in LN), and conventional p-values. There are five macro factors that appear to be statistically significant at the ten-percent level, among which three are significant at the five-percent level. The Wald statistic for H_0 far exceeds the critical values for conventional significant levels. Taken at face value, this evidence suggests that macro factors have strong predictive power, above and beyond the information contained in the yield curve.

How robust are these econometric results? We first check the warning flags. As usual, the first two yield PCs are very persistent, with autocorrelations of 0.98 and 0.94. The most persistent macro variables have first-order autocorrelations of around 0.75, so the persistence of x_{2t} is lower than in the data of JPS but still considerable. As always, the yield PCs strongly violate strict exogeneity by construction, for the reasons explained in the previous section. Based on these indicators, it appears that small-sample problems may well distort the results of conventional inference methods.

To address the potential small-sample problems we again bootstrapped 5000 data sets of artificial yields and macro data in which H_0 is true in population. The samples each contain 516 observations, which corresponds to the length of the original data sample. We report results only for the simple bootstrap without bias correction, because the bias in the VAR for x_{2t} is estimated to be small. Note that LN also considered bootstrap inference, but their main bootstrap design imposed the expectations hypothesis, in order to test whether excess

[34] We were able to closely replicate the results in LN's tables 4 through 7, and have also applied our techniques to those regressions, which led to qualitatively similar results.

returns are predictable by macro factors and the CP factor. Using this setting, LN produced convincing evidence that excess returns are predictable, which is fully consistent with our results. Our null hypothesis of interest, however, is that excess returns are predictable only by current yields. While LN also reported results for a bootstrap under the alternative hypothesis, our bootstrap under the null provides more accurate tests of the spanning hypothesis and allows us to estimate the size of conventional tests under the null (see also footnote 23).

Table 2 shows that the observed increase in predictive power from adding macro factors to the regression, measured by $\bar{R}^2$, would not be implausible if the null hypothesis were true, as the increase in $\bar{R}^2$ is within the 95% bootstrap interval. And as seen in Table 5, our bootstrap finds that only three coefficients are significant at the ten-percent level (instead of five using conventional critical values), and one at the five-percent level (instead of three). While the Wald statistic is significant even compared to the critical value from the bootstrap distribution, the evidence is weaker than when using the asymptotic distribution.

We again use the bootstrap to estimate the size and power of the different tests with a nominal size of five percent. The results, reported in Table 5, reveal that the conventional t-tests have modest size distortions, with true size of 8-14% instead of the nominal five percent. But the Wald test is seriously distorted, with a true size of 32 percent. The Wald test compounds the problems resulting from the non-standard small-sample distribution of each of the eight coefficient estimates for x_{2t}, and therefore ends up with a large size distortion. By contrast, our proposed bootstrap test has close to correct size. They also have good power, in particular the bootstrap Wald test.

Again there are several years of data that have arrived since the original LN analysis was conducted.[35] We repeated our analysis using the same 1985-2016 sample period that we used to reassess the results of JPS. There it was a strictly larger sample than the original, but here our new sample adds data at the end but leaves some out at the beginning. Reasons for interest in this sample period include the significant break in monetary policy in the early 1980s, the advantages of having a uniform sample period for comparison across all the different studies considered in our paper, and investigating robustness of the original claims in describing data since the papers were originally published. The results, shown in the right panel of Table 2 and the bottom panel of Table 5, show that over the later sample period, the evidence for the predictive power of macro factors is quite weak. The increases in $\bar{R}^2$ in Table 2 are not statistically significant, being squarely within the bootstrap intervals under the spanning hypothesis. The Wald test rejects H_0 when using asymptotic critical values, but

[35] To construct the macro factors for the 1985-2016 sample period, we used the macro data set of McCracken and Ng (2014) and transformed the data and extracted the PCs in the same way as LN did. Using the data constructed in this way, we also obtained results similar to LN's over their original sample period.

is very far from significant when using bootstrap critical values. Duffee (2013b, Section 7) has also noted problems with the stability of the results in Cochrane and Piazzesi (2005) and Ludvigson and Ng (2010) across different sample periods.

We also repeated the pure OOS exercise and report the results in the second row of Table 4. In contrast to the results for JPS (in the first row), we find that the unrestricted model which includes macro variables does better both in-sample and OOS than the model that only includes yield PCs. Adding the eight macro factors reduces the MSE for predicted returns over the 2009-2016 period by 22%. However, this improvement is not large enough to be statistically significant based on the DM test.

Overall, these results again show that conventional measures of fit and hypothesis tests are not reliable for assessing the spanning hypothesis. Furthermore, the evidence that macro factors have predictive power beyond the information already contained in yields is weaker than the results in LN would initially have suggested. Both small-sample econometric problems as well as subsample stability raise concerns about the robustness of the results.[36]

5 Trend inflation

Cieslak and Povala (2015) (henceforth CPO) presented evidence that measures of the trend in inflation can help to estimate risk premia in bond returns. They found this using a variety of measures of trend inflation. Their strongest results (and the specification we investigate here) calculates the trend in inflation using a very slowly adjusting weighted average of observed inflation rates,

$$\tau_t = (1-\nu)\textstyle\sum_{i=0}^{t-i}\nu^i\pi_{t-i}, \tag{17}$$

for π_t the month t year-over-year inflation in the core CPI and $\nu = 0.987$. CPO found that although τ_t alone does not predict excess returns, when added to a regression that also includes yields, the inflation trend becomes highly significant and the predictive power of yields improves tremendously as well.

CPO calculated standard errors using the Wei and Wright (2013) reverse regression (RR) approach as a way to mitigate the problems resulting form overlapping observations identified in Section 2.5. The RR approach uses the insight of Hodrick (1992) that it is beneficial to base inference in predictions for overlapping returns on estimates from regressions of one-period (non-overlapping) returns on cumulated predictors, and extends Hodrick's approach to perform inference about other hypotheses than the absence of predictability. We also use the RR approach throughout this section as we replicate and extend CPO's results.

[36] Appendix D reports additional results for predictive regressions with return-forecasting factors, using an empirical approach that was also advocated by LN. These results reinforce our conclusions.

To reproduce CPO's key results in a similar structure to those used in discussing the previous two studies, let y_{t+h} denote a weighted average of the annual excess returns on 2- to 15-year bonds, x_{1t} a constant and the first three PCs of yields, and $x_{2t} = \tau_t$.[37] The first three rows of Table 6 reproduce CPO's conclusion that the ability of the PCs alone to predict excess returns is modest and only the slope is a significant predictor of bond returns, consistent with the long-standing results of Campbell and Shiller (1991). But when τ_t is added (rows 4-6), the trend is highly significant, and the values and statistical significance of the coefficients on $PC1$ and $PC2$ increase tremendously as well.

The value of τ_t is plotted in Figure 2 along with the yield on a 10-year bond. Both τ_t and nominal interest rates exhibited an upward trend until the early 1980s and a distinct downward trend since then. From the start to the end of CPO's original sample, the value of τ_t fell by more than 200 basis points and the 10-year yield by over 400 basis points. The variable τ_t is also extremely persistent, with an autocorrelation of 0.9985. The analysis in Section 2.4 showed that in a setting like this, the problems from standard error bias can become much worse due to the presence of trends, and both the predictive power of τ_t and its apparent usefulness in refining the predictive power of the PCs could be spurious.

We again investigate these concerns using our bootstrap.[38] In this case, because of the very high persistence of τ_t we use the bias-corrected bootstrap. The key question is the following: For data generated under the null hypothesis that x_{1t} alone is useful in predicting returns, how often do we reject this null hypothesis? This estimate of the true size of the RR t-test is 42.5%, as reported in Table 6. The enormous size distortion results from the simultaneous presence of multiple problems, namely standard error bias, trends in x_{1t} and x_{2t}, and overlapping annual observations. We can use the bootstrap to investigate further the specific features of the DGP that lead to this poor test size.

The main problem is the presence of trends in $PC1$ and τ_t. In our bootstrap DGP x_{1t} and x_{2t} are highly persistent but stationary series, with the trend in the observed sample coming from the fact that the initial values for $PC1$ and τ_1 are the historical values in 1971, which are significantly above the population means implied by the coefficients estimated from the entire sample. When we instead initialize the bootstrap samples at the population means, so that trends are absent by construction, the size of the RR t-test is only 16% instead of 45%.

[37] We use zero-coupon yields with one to fifteen years maturity from Gürkaynak et al. (2007) and a one-month T-bill rate from the Center for Research in Security Prices (CRSP). For the dependent variable we use the same type of weighted average of excess returns as CPO, where returns are divided by the bond's duration before being averaged.

[38] Our bootstrap uses a VAR(1) for yield PCs and an AR(1) for the inflation trend. While more sophisticated bootstrap designs for inflation and the inflation trend are possible—e.g., calculating the bootstrapped inflation trend as a moving average of inflation simulated from an ARIMA model—we have found that our key results remain essentially unaffected by this choice.

This reveals that the size distortions in the CPO analysis arise primarily from the problems with trending variables analyzed in Section 2.4.[39]

The interaction of the trends present in $PC1$ and τ_t renders the inference about the coefficients on both predictors highly unreliable, as the following exercise shows. If we regress y_{t+h} on x_{1t} alone, RR t-test rejects the false null hypothesis that the coefficient on $PC1$ is zero in only 17% of the bootstrap samples. Likewise, if we regress y_{t+h} on a constant and τ_t alone, we reject the null hypothesis that $\beta_2 = 0$ in 14% of the bootstrap samples.[40] However, when both x_{1t} and x_{2t} are in the regression, we reject the hypothesis that the coefficient on $PC1$ (τ_t) is zero in 53% (45%) of the samples. The typical finding would thus be that although both $PC1$ and τ_t individually have almost no predictive power for y_{t+h}, when both are added to the same regression they typically appear statistically significant. We have also found that adding τ_t to the regression doubles the root-mean-square error for the coefficient estimate on $PC1$ around its true value. This suggests that rather than helping refine the predictive power of x_{1t}, the addition of x_{2t} in fact leads to a substantial deterioration of the forecasting model. The reason for all of these problems is the simultaneous presence of trends in x_{1t} and τ_t which substantially distorts the inference about the spanning hypothesis, in line with the econometric theory in Section 2.4.

Furthermore, the problems stemming from overlapping observations are only partially alleviated by the RR test. If instead of RR standard errors we use Newey-West standard errors with the usual 18 lags, the test of $\beta_2 = 0$ has an even larger size of 56%, compared to the 45% size of the RR test, so the RR approach helps some. However, in the absence of overlapping observations, a t-test of the same hypothesis has a size of 34%. Since this is quite a bit below 45% the RR test apparently does not completely solve the problem of overlapping observations.[41]

Notwithstanding, we emphasize that these concerns cannot entirely explain the size of the effects found by CPO. As seen in Table 2, it would not be surprising to see the estimated $\bar{R}^2$ go

[39] By contrast, in the JPS data we found that the biggest single source of the size distortions is the use of overlapping returns and Newey-West standard errors. And in the LN data it is a combination of the overlapping returns and the presence of a relatively large number of predictors in x_{2t}, which magnifies the size distortions.

[40] In our DGP, as in the data, bond risk premia are driven mainly by $PC2$; the population coefficient on $PC1$ is nonzero but close to zero. And τ_t is correlated with x_{1t} in the bootstrap DGP so τ_t by itself also has some predictive power. But for both $PC1$ and τ_t the predictive power is usually not big enough for the RR t-test to detect.

[41] To obtain this result we set $h = 1$ and calculate y_{t+1} as monthly excess returns, using the usual approximation $i_{n-1,t+1} \approx i_{n,t+1}$ and the one-month T-bill rate. The t-test in this case uses White's heteroskedasticity robust standard errors (as in, for example, Duffee, 2013b, Section 7). It is well-known that RR standard errors, just like Hodrick's standard errors, do not eliminate the problem of Stambaugh bias; note for example the size distortions in Table 1 of Wei and Wright (2013). Therefore it is unsurprising that this approach does not eliminate standard error bias.

from 15% when only PCs are used to as high as 40% when τ_t is added. In the data, however, we find an $\bar{R}^2$ of 46% when τ_t is added, too big to be attributed to the factors captured by our bootstrap alone. We also see from Table 6 that while it would not be surprising to see the RR t-statistic for β_2 as high as 3.6; the value observed in the data of 6.2 is much too big to be explained by these considerations alone.[42]

We also reestimated the predictive regressions over the 1985-2016 period that we have used as a common comparison with the other studies. The increase in $\bar{R}^2$ is smaller and no longer statistically significant on this dataset, as seen in the last three columns of Table 2. Compared to the bootstrap small-sample critical values, the RR t-statistic is only marginally statistically significant at the 5% level, as seen in the second panel of Table 6. Note that in the post-1985 sample the downward drift in the level of yields and the inflation trend is even more pronounced, adding to the econometric problems caused by trends in explanatory variables.

We also evaluated the true OOS usefulness of τ_t using new data after the end of CPO's sample period, which we report in line 3 of Table 4. Whereas within CPO's original sample the trend reduces the MSE by 40%, for the data that have arrived since 2011 including the inflation trend actually increases the MSE by 221%, and based on the DM statistic this deterioration is strongly statistically significant.

In sum, there are two possibly complementary explanations for the strong in-sample predictive power of the inflation trend τ_t documented by CPO. First, the addition of τ_t may truly help improve forecasts of bond returns, for example because accounting for the common, slow-moving trend in yields and τ_t might uncover additional predictive power. This interpretation is supported by the fact that CPO's finding survives our bootstrap correction for small-sample problems, at least in their original data set. But a second explanation, suggested by the theoretical arguments in Section 2.4, is that a substantial portion of the apparent incremental predictive power of τ_t arises spuriously from the presence of trends. This second explanation is supported by our bootstrap analysis of the the role of trends, by the results in the 1985-2016 sample, and by the poor out-of sample performance of the model that includes τ_t vis-a-vis a model that imposes the spanning hypothesis. Clearly one needs to exercise particularly great care in interpreting evidence against the spanning hypothesis in a situation with trending predictors.

[42] We note from Figure 2 that τ_t is even better characterized as exhibiting two different trends rather than a single downward trend as captured in our bootstrap. We have found in simulations that such breaking trends can substantially exacerbate the problems that arise from a single trend.

6 Higher-order PCs of yields

Cochrane and Piazzesi (2005) (henceforth CP) documented several striking facts about excess bond returns. They showed that a tent-shaped combination of forward rates predicts annual excess returns on different long-term bonds with an R^2 of up to 37% (and even up to 44% when lags are included). Importantly for our context, CP found that the first three PCs of yields—level, slope, and curvature—did not fully capture this predictability, but that the fourth and fifth PC were also very helpful. As usual, the first three PCs explain a large share of the cross-section variation in yields (99.97% in their data), but CP found that the other two PCs, which explain only 0.03% of the cross-section variation in yields, are statistically important for predicting excess bond returns. In particular, the fourth PC appeared "very important for explaining expected returns" (p. 147). Here we assess the robustness of this finding, by revisiting the null hypothesis that only the first three PCs, but not higher-order PCs, predict excess returns.

The last panel of Table 2 shows (unadjusted) R^2 for predictive regressions for the average excess bond return using three and five PCs as predictors, and the first entries replicate the results of CP. In Table 7 we report the results of HAC inference for the regressions with 5 PCs using Newey-West standard errors with 18 lags, and the Wald statistic is identical to that reported by CP in their Table 4. The p-values indicate that $PC4$ is very strongly statistically significant, and that the spanning hypothesis would be rejected.

We then use our bootstrap procedure to obtain robust inference about the relevance of the predictors PC4 and PC5. We find that CP's result is not due to small-sample size distortions. The persistence of higher-order PCs is quite low, so that the size distortions of conventional tests are small. And the Newey-West t-statistic on $PC4$ is far too large to be accounted for by the kinds of factors identified in Section 2. Likewise the increase in R^2 reported by CP would be quite implausible under the null hypothesis, as it falls far outside the 95% bootstrap interval under the null.

In the last three columns of Table 2 and the bottom panel of Table 7 we report results for the 1985–2016 sample period. In this case, the increase in R^2 due to inclusion of higher-order PCs is comfortably inside the 95% bootstrap intervals, and the coefficients on $PC4$ and $PC5$ are not significant for any method of inference.[43]

CP's sample period ended more than ten years prior to the time of this writing, giving us the longest true OOS period among the studies considered. The last row of Table 4 shows that in contrast to the in-sample estimates, where including $PC4$ and $PC5$ reduces the MSE

[43] Consistent with this finding, an influence analysis of the predictive power of $PC4$ in the full sample indicates that the observations with the largest leverage and influence are almost all clustered in the early and mid 1980s.

by 11%, OOS predictions become less accurate, with the MSE increased by 21%, when the null hypothesis is not imposed. While the DM test does not reject the hypothesis that both models have equal predictive accuracy in population, restricting the predictive model to use only the level, slope and curvature leads to more stable and more accurate return predictions in post-publication data.

It is worth emphasizing the similarities and differences between CP's results and ours. Their central claim, with which we concur, is that the factor they have identified is a useful and stable predictor of bond returns. CP conducted tests of the usefulness of their return-forecasting factor for predicting returns across different subsamples, a result that we have been able to reproduce and confirm. But their factor is a function of all 5 PC's, and our results suggest that it is mainly $PC1$-$PC3$, and not the addition of $PC4$ and $PC5$, that makes this factor a robust predictor of bond returns. We have shown that these higher-order PCs are insignificant in the 1985–2016 sample and in true out-of-sample forecasting. In additional, unreported results we found the same results for most of the subsample periods that CP considered, as well as for the 1952-2010 sample period considered by Duffee (2013b, Section 7). We conclude that the predictive power of higher-order factors is tenuous and sample-dependent, and that there is no compelling evidence that the first three PCs of yields are insufficient to estimate bond risk premia.[44]

7 Other studies

Several other studies have also reported evidence that might appear to be inconsistent with the spanning hypothesis. Cooper and Priestley (2008) concluded that the output gap contains useful information for forecasting interest rates, while Greenwood and Vayanos (2014) found the same for measures of Treasury bond supply. We have repeated our analysis using the datasets in these studies and found that evidence against the spanning hypothesis in these two cases is even weaker than for any of the studies discussed in Sections 3 to 6. Details of our investigations are reported in Appendices E and F.

8 Conclusion

Conventional tests of whether variables other than the level, slope and curvature can help predict bond returns have significant size distortions, and the R^2 of the regression can in-

[44] Cattaneo and Crump (2014) also investigated the robustness of the results of Cochrane and Piazzesi (2005) and obtained even more negative results: Using a new HAC test proposed by Müller (2014) they did not reject the null hypothesis that the CP factor had no predictive power in a variety of in-sample and OOS specifications.

crease dramatically when other variables are added to the regression even if they have no true explanatory power. We proposed a simple parametric bootstrap to carry out inference that is robust to the resulting small-sample problems. We used this bootstrap approach to reexamine the usefulness of proposed predictors in six widely cited studies, in both the original data and in a common sample period that includes newly available data. In addition, we calculated true out-of-sample forecasts. Our overall finding is that conventional tests are highly unreliable, and that as a result the evidence that variables other than the current level, slope and curvature predict excess bond returns is substantially less convincing than the original research would have led us to believe.

References

Adrian, Tobias, Richard K. Crump, and Emanuel Moench (2013) "Pricing the Term Structure with Linear Regressions," *Journal of Financial Economics*, Vol. 110, pp. 110–138.

Andrews, Donald W. K. (1991) "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation," *Econometrica*, Vol. 59, pp. 817–858.

Ang, Andrew and Geert Bekaert (2007) "Stock return predictability: Is it there?" *Review of Financial studies*, Vol. 20, pp. 651–707.

Bansal, Ravi and Ivan Shaliastovich (2013) "A Long-Run Risks Explanation of Predictability Puzzles in Bond and Currency Markets," *Review of Financial Studies*, Vol. 26, pp. 1–33.

Basawa, Ishwar V, Asok K Mallik, William P McCormick, Jaxk H Reeves, and Robert L Taylor (1991) "Bootstrapping unstable first-order autoregressive processes," *Annals of Statistics*, pp. 1098–1101.

Bauer, Michael D. and Glenn D. Rudebusch (2017) "Resolving the Spanning Puzzle in Macro-Finance Term Structure Models," *Review of Finance*, Vol. 21, pp. 511–553.

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

Bekaert, Geert, Eric Engstrom, and Yuhang Xing (2009) "Risk, Uncertainty, and Asset Prices," *Journal of Financial Economics*, Vol. 91, pp. 59–82.

Bekaert, Geert and Robert J Hodrick (2001) "Expectations hypotheses tests," *The journal of finance*, Vol. 56, pp. 1357–1394.

Berkowitz, Jeremy and Lutz Kilian (2000) "Recent developments in bootstrapping time series," *Econometric Reviews*, Vol. 19, pp. 1–48.

Bikbov, Ruslan and Mikhail Chernov (2010) "No-Arbitrage Macroeconomic Determinants of the Yield Curve," *Journal of Econometrics*, Vol. 159, pp. 166–182.

Campbell, John Y. and John H. Cochrane (1999) "By force of habit: A consumption-based explanation of aggregate stock market behavior," *Journal of Political Economy*, Vol. 107, pp. 205–251.

Campbell, John Y. and Robert J. Shiller (1991) "Yield Spreads and Interest Rate Movements: A Bird's Eye View," *Review of Economic Studies*, Vol. 58, pp. 495–514.

Campbell, John Y and Motohiro Yogo (2006) "Efficient tests of stock return predictability," *Journal of financial economics*, Vol. 81, pp. 27–60.

Cattaneo, Matias D. and Richard K. Crump (2014) "Comment," *Journal of Business & Economic Statistics*, Vol. 32, pp. 324–329.

Cavanagh, Christopher L, Graham Elliott, and James H Stock (1995) "Inference in Models with Nearly Integrated Regressors," *Econometric theory*, Vol. 11, pp. 1131–1147.

Chan, Ngai Hang (1988) "The parameter inference for nearly nonstationary time series," *Journal of the American Statistical Association*, Vol. 83, pp. 857–862.

Cieslak, Anna and Pavol Povala (2015) "Expected Returns in Treasury Bonds," *Review of Financial Studies*, Vol. 28, pp. 2859–2901.

Cochrane, John H. and Monika Piazzesi (2005) "Bond Risk Premia," *American Economic Review*, Vol. 95, pp. 138–160.

Cooper, Ilan and Richard Priestley (2008) "Time-Varying Risk Premiums and the Output Gap," *Review of Financial Studies*, Vol. 22, pp. 2801–2833.

Dewachter, Hans and Marco Lyrio (2006) "Macro Factors and the Term Structure of Interest Rates," *Journal of Money, Credit and Banking*, Vol. 38, pp. 119–140.

Diebold, Francis X. and Robert S. Mariano (1995) "Comparing Predictive Accuracy," *Journal of Business & economic statistics*, Vol. 13, pp. 253–263.

Diebold, Francis X., Glenn D. Rudebusch, and S. Boragan Aruoba (2006) "The Macroeconomy and the Yield Curve: A Dynamic Latent Factor Approach," *Journal of Econometrics*, Vol. 131, pp. 309–338.

Duffee, Gregory R. (2011a) "Forecasting with the Term Structure: the Role of No-Arbitrage," Working Paper January, Johns Hopkins University.

——— (2011b) "Information In (and Not In) the Term Structure," *Review of Financial Studies*, Vol. 24, pp. 2895–2934.

——— (2013a) "Bond Pricing and the Macroeconomy," in Milton Harris George M. Constantinides and Rene M. Stulz eds. *Handbook of the Economics of Finance*, Vol. 2, Part B: Elsevier, pp. 907–967.

——— (2013b) "Forecasting Interest Rates," in Graham Elliott and Allan Timmermann eds. *Handbook of Economic Forecasting*, Vol. 2, Part A: Elsevier, pp. 385–426.

Fama, Eugene F. and Robert R. Bliss (1987) "The Information in Long-Maturity Forward Rates," *The American Economic Review*, Vol. 77, pp. 680–692.

Ferson, Wayne E, Sergei Sarkissian, and Timothy T Simin (2003) "Spurious Regressions in Financial Economics?" *Journal of Finance*, Vol. 58, pp. 1393–1414.

Giacoletti, Marco, Kristoffer T. Laursen, and Kenneth J. Singleton (2016) "Learning, Dispersion of Beliefs, and Risk Premiums in an Arbitrage-free Term Structure Model," unpublished manuscript.

Goetzmann, William N and Philippe Jorion (1993) "Testing the predictive power of dividend yields," *The Journal of Finance*, Vol. 48, pp. 663–679.

Greenwood, Robin and Dimitri Vayanos (2014) "Bond Supply and Excess Bond Returns," *Review of Financial Studies*, Vol. 27, pp. 663–713.

Gürkaynak, Refet S., Brian Sack, and Jonathan H. Wright (2007) "The U.S. Treasury yield curve: 1961 to the present," *Journal of Monetary Economics*, Vol. 54, pp. 2291–2304.

Gürkaynak, Refet S. and Jonathan H. Wright (2012) "Macroeconomics and the Term Structure," *Journal of Economic Literature*, Vol. 50, pp. 331–367.

Hall, Peter and Susan R. Wilson (1991) "Two Guidelines for Bootstrap Hypothesis Testing," *Biometrics*, Vol. 47, pp. 757–762.

Hamilton, James D. (1994) *Time Series Analysis*: Princeton University Press.

Hamilton, James D. and Jing Cynthia Wu (2012) "Identification and estimation of Gaussian affine term structure models," *Journal of Econometrics*, Vol. 168, pp. 315–331.

——— (2014) "Testable Implications of Affine Term Structure Models," *Journal of Econometrics*, Vol. 178, pp. 231–242.

Hodrick, Robert J (1992) "Dividend yields and expected stock returns: Alternative procedures for inference and measurement," *Review of Financial studies*, Vol. 5, pp. 357–386.

Hördahl, Peter, Oreste Tristani, and David Vestin (2006) "A Joint Econometric Model of Macroeconomic and Term-Structure Dynamics," *Journal of Econometrics*, Vol. 131, pp. 405–444.

Horowitz, Joel L. (2001) "The Bootstrap," in J.J. Heckman and E.E. Leamer eds. *Handbook of Econometrics*, Vol. 5: Elsevier, Chap. 52, pp. 3159–3228.

Ibragimov, Rustam and Ulrich K. Müller (2010) "t-Statistic Based Correlation and Heterogeneity Robust Inference," *Journal of Business and Economic Statistics*, Vol. 28, pp. 453–468.

Joslin, Scott, Marcel Priebsch, and Kenneth J. Singleton (2014) "Risk Premiums in Dynamic Term Structure Models with Unspanned Macro Risks," *Journal of Finance*, Vol. 69, pp. 1197–1233.

Kendall, M. G. (1954) "A note on bias in the estimation of autocorrelation," *Biometrika*, Vol. 41, pp. 403–404.

Kilian, Lutz (1998) "Small-sample confidence intervals for impulse response functions," *Review of Economics and Statistics*, Vol. 80, pp. 218–230.

Le, Anh and Kenneth J. Singleton (2013) "The Structure of Risks in Equilibrium Affine Models of Bond Yields," May, unpublished manuscript.

Lewellen, Jonathan, Stefan Nagel, and Jay Shanken (2010) "A skeptical appraisal of asset pricing tests," *Journal of Financial Economics*, Vol. 96, pp. 175–194.

Litterman, Robert and J. Scheinkman (1991) "Common Factors Affecting Bond Returns," *Journal of Fixed Income*, Vol. 1, pp. 54–61.

Ludvigson, Sydney C. and Serena Ng (2009) "Macro Factors in Bond Risk Premia," *Review of Financial Studies*, Vol. 22, pp. 5027–5067.

Ludvigson, Sydney C and Serena Ng (2010) "A Factor Analysis of Bond Risk Premia," *Handbook of Empirical Economics and Finance*, p. 313.

Mankiw, N. Gregory and Matthew D. Shapiro (1986) "Do we reject too often? Small sample properties of tests of rational expectations models," *Economics Letters*, Vol. 20, pp. 139–145.

McCracken, Michael W. and Serena Ng (2014) "FRED-MD: A Monthly Database for Macroeconomic Research," working paper, Federal Reserve Bank of St. Louis.

Müller, Ulrich K. (2014) "HAC Corrections for Strongly Autocorrelated Time Series," *Journal of Business and Economic Statistics*, Vol. 32.

Nabeya, Seiji and Bent E Sørensen (1994) "Asymptotic distributions of the least-squares estimators and test statistics in the near unit root model with non-zero initial value and local drift and trend," *Econometric Theory*, Vol. 10, pp. 937–966.

Nelson, Charles R and Myung J Kim (1993) "Predictable stock returns: The role of small sample bias," *The Journal of Finance*, Vol. 48, pp. 641–661.

Newey, Whitney K and Kenneth D West (1987) "A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, Vol. 55, pp. 703–08.

Phillips, Peter CB (1988) "Regression theory for near-integrated time series," *Econometrica: Journal of the Econometric Society*, pp. 1021–1043.

Piazzesi, Monika, Juliana Salomao, and Martin Schneider (2015) "Trend and Cycle in Bond Premia," March, unpublished manuscript.

Piazzesi, Monika and Martin Schneider (2007) "Equilibrium Yield Curves," in *NBER Macroeconomics Annual 2006, Volume 21*: MIT Press, pp. 389–472.

Pope, Alun L. (1990) "Biases of Estimators in Multivariate Non-Gaussian Autoregressions," *Journal of Time Series Analysis*, Vol. 11, pp. 249–258.

Rudebusch, Glenn D. and Eric T. Swanson (2012) "The bond premium in a DSGE Model with Long-Run Real and Nominal Risks," *American Economic Journal: Macroeconomics*, Vol. 4, pp. 105–143.

Rudebusch, Glenn D. and Tao Wu (2008) "A Macro-Finance Model of the Term Structure, Monetary Policy, and the Economy," *Economic Journal*, Vol. 118, pp. 906–926.

Stambaugh, Robert F. (1999) "Predictive regressions," *Journal of Financial Economics*, Vol. 54, pp. 375–421.

Stock, James H (1991) "Confidence intervals for the largest autoregressive root in US macroeconomic time series," *Journal of Monetary Economics*, Vol. 28, pp. 435–459.

Stock, James H. (1994) "Unit roots, structural breaks and trends," in Robert F. Engle and Daniel L. McFadden eds. *Handbook of Econometrics*, Vol. 4: Elsevier, Chap. 46, pp. 2739–2841.

Wachter, Jessica A. (2006) "A Consumption-Based Model of the Term Structure of Interest Rates," *Journal of Financial Economics*, Vol. 79, pp. 365–399.

Wei, Min and Jonathan H Wright (2013) "Reverse Regressions And Long-Horizon Forecasting," *Journal of Applied Econometrics*, Vol. 28, pp. 353–371.

Welch, Ivo and Amit Goyal (2008) "A comprehensive look at the empirical performance of equity premium prediction," *Review of Financial Studies*, Vol. 21, pp. 1455–1508.

Table 1: Simulation study of standard error bias

ρ	δ	Coefficient bias		SE bias	Size		
		β_1	β_2	(%)	simulated	asymptotic	bootstrap
$\mu_1 = \mu_2 = 0$							
0.99	0.0	0.000	0.000	-4.7	0.050	0.047	0.048
0.00	1.0	-0.010	0.000	-0.6	0.050	0.051	0.050
0.90	1.0	-0.052	0.000	-15.4	0.085	0.086	0.057
0.99	0.8	-0.055	0.000	-23.2	0.113	0.112	0.072
0.99	1.0	-0.068	0.000	-29.8	0.151	0.151	0.082
$\mu_1 = 0, \mu_2 = 1$							
0.99	0.0	0.000	0.000	-5.1	0.050		0.049
0.00	1.0	-0.010	0.000	-0.5	0.050		0.050
0.90	1.0	-0.053	0.000	-17.1	0.089		0.057
0.99	0.8	-0.071	0.000	-42.4	0.183		0.077
0.99	1.0	-0.088	0.000	-50.8	0.268		0.085
$\mu_1 = 1, \mu_2 = 1$							
0.99	0.0	0.000	0.000	-4.0	0.050		0.047
0.00	1.0	-0.010	0.000	-0.5	0.050		0.050
0.90	1.0	-0.037	0.017	-12.0	0.081		0.054
0.99	0.8	-0.036	0.035	-12.1	0.168		0.056
0.99	1.0	-0.045	0.044	-16.0	0.241		0.058

Coefficient bias, standard error bias, and test size in simulation study for predictive regressions $y_{t+1} = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + u_{t+1}$ in sample of size $T = 100$ from data-generating process (DGP) with x_{1t} and x_{2t} following AR(1) processes, $\beta_0 = \beta_1 = \beta_2 = 0$, and different values of $\rho_1 = \rho_2 = \rho$ and δ. For details on the DGP refer to text. The coefficient bias is reported as $E(\hat{\beta}_i) - \beta_i$. The standard error bias is reported as $E[(\hat{\sigma}_{\hat{\beta}_2}) - \sigma_{\hat{\beta}_2}]/\sigma_{\hat{\beta}_2}$. The last three columns report the size (i.e., frequency of rejections) of tests of $H_0 : \beta_2 = 0$ with a nominal size of five percent, for a conventional t-test—according to both regressions in simulated small samples and the local-to-unity asymptotic distribution—and for the bootstrap test.

Table 2: In-sample predictive power in excess-return regressions

	$\bar{R}^2_1$	$\bar{R}^2_2$	$\bar{R}^2_2 - \bar{R}^2_1$	$\bar{R}^2_1$	$\bar{R}^2_2$	$\bar{R}^2_2 - \bar{R}^2_1$
JPS	Original sample: 1985–2008			Later sample: 1985–2016		
Data	0.19	0.38	0.19	0.17	0.21	0.04
Bootstrap	0.32	0.38	0.06	0.28	0.32	0.05
	(0.11, 0.55)	(0.15, 0.60)	(0.00, 0.20)	(0.08, 0.49)	(0.12, 0.53)	(0.00, 0.17)
BC bootstrap	0.35	0.41	0.06	0.28	0.33	0.05
	(0.08, 0.62)	(0.13, 0.67)	(0.00, 0.23)	(0.06, 0.52)	(0.11, 0.57)	(0.00, 0.20)
Ludvigson-Ng	Original sample: 1964–2007			Later sample: 1985–2016		
Data	0.25	0.35	0.10	0.14	0.24	0.10
Bootstrap	0.21	0.24	0.03	0.29	0.34	0.05
	(0.05, 0.38)	(0.08, 0.42)	(0.00, 0.11)	(0.09, 0.51)	(0.13, 0.55)	(0.00, 0.16)
Cieslak-Povala	Original sample: 1971–2011			Later sample: 1985–2016		
Data	0.12	0.46	0.34	0.16	0.35	0.19
BC bootstrap	0.15	0.22	0.07	0.27	0.34	0.07
	(0.02, 0.34)	(0.06, 0.40)	(0.00, 0.21)	(0.05, 0.53)	(0.10, 0.57)	(0.00, 0.23)
Cochrane-Piazzesi	Original sample: 1964–2003			Later sample: 1985–2016		
Data	0.26	0.35	0.09	0.15	0.18	0.03
Bootstrap	0.21	0.22	0.01	0.30	0.31	0.01
	(0.05, 0.39)	(0.06, 0.40)	(0.00, 0.02)	(0.10, 0.51)	(0.11, 0.52)	(0.00, 0.05)

Adjusted $\bar{R}^2$ for regressions of annual excess bond returns on three PCs of the yield curve ($\bar{R}^2_1$) and on three yield PCs together with the additional proposed predictors x_{2t}, well as the difference in adjusted $\bar{R}^2$. The additional predictors, which are described in more detail in the text, are: for JPS, measures of growth and inflation; for Ludvigson-Ng, eight PCs of a large set of macro variables; for Cieslak-Povala, a moving-average estimate of the inflation trend; and for Cochrane-Piazzesi, the fourth and fifth PC of yields. The results in the left half of the table are for the original sample period in each paper; the right half of the table is for the 1985–2016 sample period. The excess bond return is an average across bond maturities: for JPS, from two to ten years; for Ludvigson-Ng, from two to five years; for Cieslak-Povala, from two to ten years (a weighted average); and for Cochrane-Piazzesi, from two to five years. The first row of each panel reports the values of the statistics in the original data. The following rows report bootstrap mean and 95%-quantiles (in parentheses). The bootstrap, which is described in the text, imposes the null hypothesis that x_{2t} has no incremental predictive power. For Cochrane-Piazzesi, the results are for the unadjusted R^2.

Table 3: Joslin-Priebsch-Singleton: statistical inference in excess-return regressions

	$PC1$	$PC2$	$PC3$	GRO	INF	Wald
Original sample: 1985–2008						
Coefficient	1.090	1.793	2.874	-2.200	-6.052	
HAC statistic	5.587	3.933	0.799	-2.475	-4.265	25.152
HAC p-value	**0.000**	**0.000**	0.425	**0.014**	**0.000**	**0.000**
Bootstrap 5% c.v.				3.177	3.870	22.705
Bootstrap p-value				0.109	**0.034**	**0.038**
BC bootstrap 5% c.v.				3.245	4.261	25.796
BC bootstrap p-value				0.123	0.050	0.053
Size						
HAC				0.189	0.274	0.356
Bootstrap				0.059	0.060	0.063
Power						
Bootstrap				0.204	0.904	0.893
Later sample: 1985–2016						
Coefficient	0.371	1.741	1.542	-0.429	-2.420	
HAC statistic	2.302	3.324	0.611	-0.537	-1.798	3.350
HAC p-value	**0.022**	**0.001**	0.542	0.592	0.073	0.187
Simple bootstrap 5% c.v.				3.008	3.409	18.510
Simple bootstrap p-value				0.706	0.283	0.504
BC bootstrap 5% c.v.				3.075	3.794	21.337
BC bootstrap p-value				0.713	0.317	0.552

Predictive regressions for annual excess bond returns, averaged over two- through ten-year bond maturities, using yield PCs and two macro variables that are described in the text. Results in the top panel are for the same sample period used in Joslin et al. (2014); the data used for the bottom panel is extended to December 2016. HAC statistics and p-values are calculated using Newey-West standard errors with 18 lags. The column "Wald" reports results for the χ^2 test that GRO and INF have no predictive power; the other columns report results for individual t-tests. We obtain bootstrap distributions of the test statistics under the null hypothesis that GRO and INF have no predictive power—the text describes the design of the simple and bias-corrected (BC) bootstraps. Critical values (c.v.'s) are the 95th percentile of the bootstrap distribution of the test statistics, and p-values are the frequency of bootstrap replications in which the test statistics are at least as large as in the data. Under *Size* we report estimates of the size of the tests, based on simulations from the simple bootstrap under the null hypothesis. Under *Power* we report power estimates using a bootstrap under the alternative hypothesis, as described in the text. p-values below 5% are emphasized with bold face.

Table 4: In-sample vs. out-of-sample predictive power

	In-sample			Out-of-sample			
	R_1^2	R_2^2	MSE-ratio	Start	N	MSE-ratio	DM p-value
Joslin-Priebsch-Singleton	0.191	0.381	0.759	2008:1	96	2.156	**0.005**
Ludvigson-Ng	0.258	0.359	0.850	2007:1	108	0.778	0.314
Cieslak-Povala	0.165	0.496	0.603	2011:1	60	3.213	**0.006**
Cochrane-Piazzesi	0.267	0.344	0.891	2003:1	156	1.213	0.103

In-sample vs. out-of-sample (OOS) predictive power for excess bond returns (averaged across maturities) of a restricted model with three PCs and an unrestricted model with additional predictors as suggested in each of four published studies. The in-sample period is the original sample period used in each study. The OOS period starts after the end of the in-sample period and ends in December 2016. OOS forecasts are generated using a recursive estimation scheme. N indicates the number of OOS observations. The columns also show in-sample R^2 for the restricted and unrestricted model, the in-sample ratio of mean-squared-errors (MSE) for the unrestricted relative to the restricted model, and the OOS MSE ratio, as well as the p-value of the Diebold-Mariano (DM) test for equal forecast accuracy.

Table 5: Ludvigson-Ng: statistical inference in excess-return regressions

	$PC1$	$PC2$	$PC3$	$F1$	$F2$	$F3$	$F4$	$F5$	$F6$	$F7$	$F8$	Wald
Original sample: 1964–2007												
Coefficient	0.136	2.052	-5.010	0.742	0.147	0.072	-0.528	-0.321	0.576	0.401	0.551	
HAC statistic	1.553	2.594	-2.724	1.855	0.380	0.608	-1.912	-1.307	2.221	2.361	3.036	42.073
HAC p-value	0.121	**0.010**	**0.007**	0.064	0.704	0.544	0.056	0.192	**0.027**	**0.019**	**0.003**	**0.000**
Bootstrap 5% c.v.				2.551	2.502	2.227	2.527	2.425	2.607	2.459	2.326	29.323
Bootstrap p-value				0.147	0.753	0.579	0.132	0.282	0.095	0.061	**0.012**	**0.010**
Size												
HAC				0.127	0.113	0.082	0.124	0.110	0.136	0.121	0.094	0.325
Bootstrap				0.052	0.049	0.047	0.049	0.045	0.049	0.050	0.052	0.058
Power												
Bootstrap				0.441	0.072	0.087	0.410	0.207	0.348	0.503	0.773	0.951
Later sample: 1985–2016												
Coefficient	0.207	1.492	-1.339	0.960	0.484	0.316	-0.245	-0.037	-0.158	0.230	-0.236	
HAC statistic	2.604	1.559	-0.434	2.459	1.856	0.908	-1.171	-0.136	-1.297	0.689	-0.922	22.768
HAC p-value	**0.010**	0.120	0.665	**0.014**	0.064	0.364	0.242	0.892	0.195	0.491	0.357	**0.004**
Bootstrap 5% c.v.				2.989	3.156	3.171	2.825	2.825	2.407	2.798	2.522	41.058
Bootstrap p-value				0.100	0.223	0.542	0.399	0.923	0.283	0.610	0.472	0.258

Predictive regressions for annual excess bond returns, averaged over two- through five-year bond maturities, using yield PCs and factors from a large data set of macro variables, as in Ludvigson and Ng (2010). The top panel shows the results for the original data set used by Ludvigson and Ng (2010); the bottom panel uses a data sample that starts in 1985 and ends in 2016. The bootstrap is a simple bootstrap without bias correction. For a description of the statistics in each row, see the notes to Table 3. p-values below 5% are emphasized with bold face.

Table 6: Cieslak-Povala: statistical inference in excess-return regressions

	$PC1$	$PC2$	$PC3$	τ
Original sample: 1971–2011				
Only yield PCs				
Coefficient	0.003	0.240	-0.127	
RR t-statistic	0.448	2.497	-0.630	
RR p-value	0.654	**0.013**	0.529	
Yield PCs plus inflation trend				
Coefficient	0.160	0.429	-0.059	-0.962
RR t-statistic	5.173	5.227	-0.322	-6.329
RR p-value	**0.000**	**0.000**	0.748	**0.000**
Bootstrap RR 5% c.v.				3.538
Bootstrap RR p-value				**0.000**
Size				
RR				0.425
Bootstrap				0.075
Power				
Bootstrap				0.978
Later sample: 1985–2016				
Only yield PCs				
Coefficient	0.019	0.180	-0.056	
RR t-statistic	1.825	1.639	-0.027	
RR p-value	0.069	0.102	0.978	
Yield PCs plus inflation trend				
Coefficient	0.106	0.297	0.061	-0.607
RR t-statistic	4.395	3.548	0.611	-3.708
RR p-value	**0.000**	**0.000**	0.541	**0.000**
Bootstrap RR 5% c.v.				3.580
Bootstrap RR p-value				**0.039**

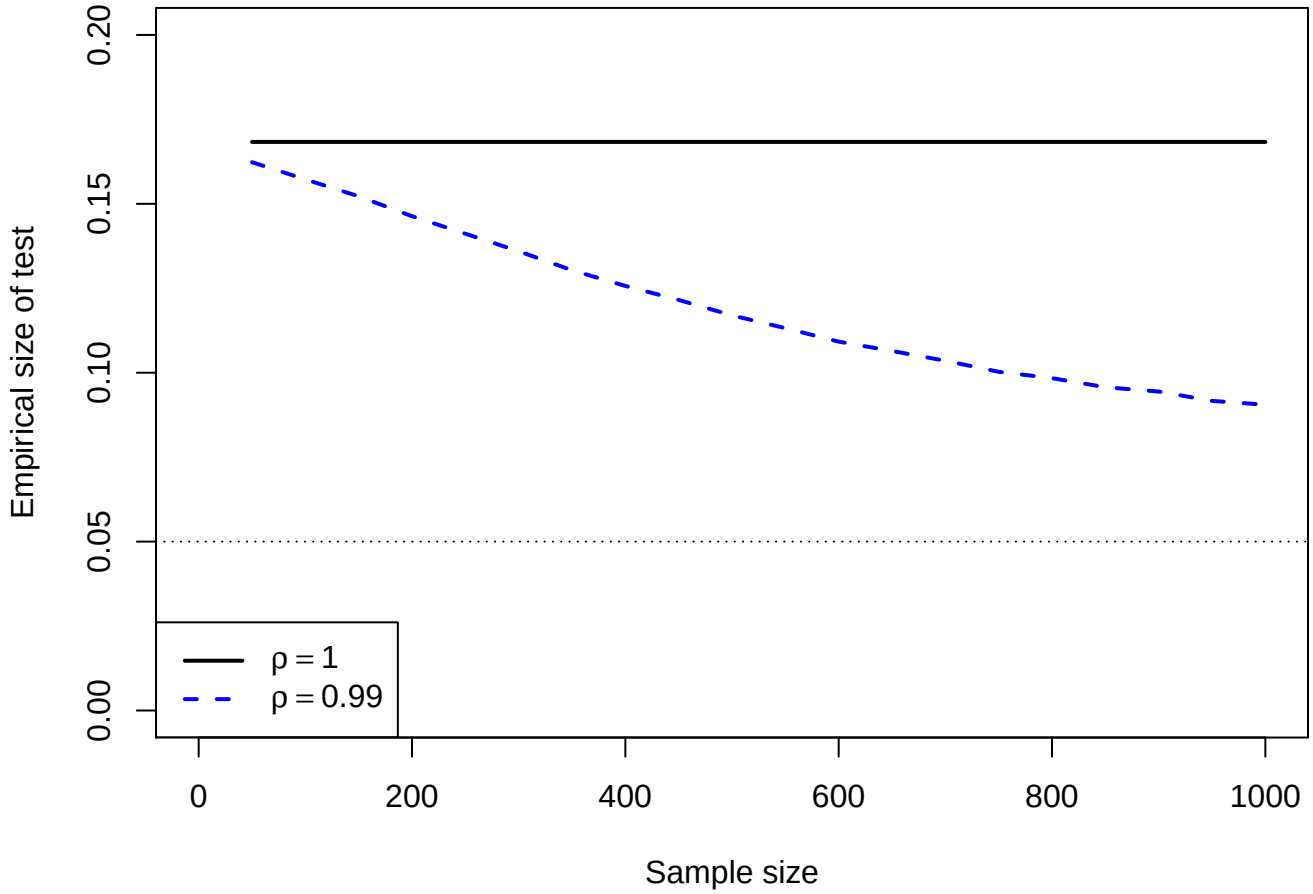
Predictive regressions for annual excess bond returns (weighted average over two- through ten-year bond maturities) using yield PCs and the moving-average estimate of inflation trend defined in equation (17). The data used for the top panel covers the same sample period as in Cieslak and Povala (2015); the data used for the bottom panel starts in 1985 and ends in 2016. Reverse regression (RR) statistics and p-values are calculated using the reverse regression delta method of Wei and Wright (2013). We obtain bootstrap distributions of the test statistics under the null hypothesis that only PCs have predictive power, in order to calculate bootstrap critical values and p-values, and to estimate the size of tests. Under *Size* we report estimates of the size of the tests based on the bootstrap samples. Under *Power* we report power estimates using a bootstrap under the alternative hypothesis, as described in the text. p-values below 5% are emphasized with bold face.

Table 7: Cochrane-Piazzesi: statistical inference in excess-return regressions

	$PC1$	$PC2$	$PC3$	$PC4$	$PC5$	Wald
Original sample: 1964–2003						
Data	0.127	2.740	-6.307	-16.128	-2.038	
HAC statistic	1.724	5.205	2.950	5.626	0.748	31.919
HAC p-value	0.085	**0.000**	**0.003**	**0.000**	0.455	**0.000**
Bootstrap 5% c.v.				2.330	2.178	8.359
Bootstrap p-value				**0.000**	0.501	**0.000**
Size						
HAC				0.091	0.077	0.112
Bootstrap				0.060	0.040	0.050
Power						
Bootstrap				0.995	0.113	0.988
Later sample: 1985–2016						
Coefficient	0.106	1.589	3.157	-9.585	-9.360	
HAC statistic	1.982	2.254	0.950	-1.460	-1.263	4.180
HAC p-value	**0.048**	**0.025**	0.343	0.145	0.207	0.124
Bootstrap 5% c.v.				2.480	2.445	9.962
Bootstrap p-value				0.239	0.295	0.264

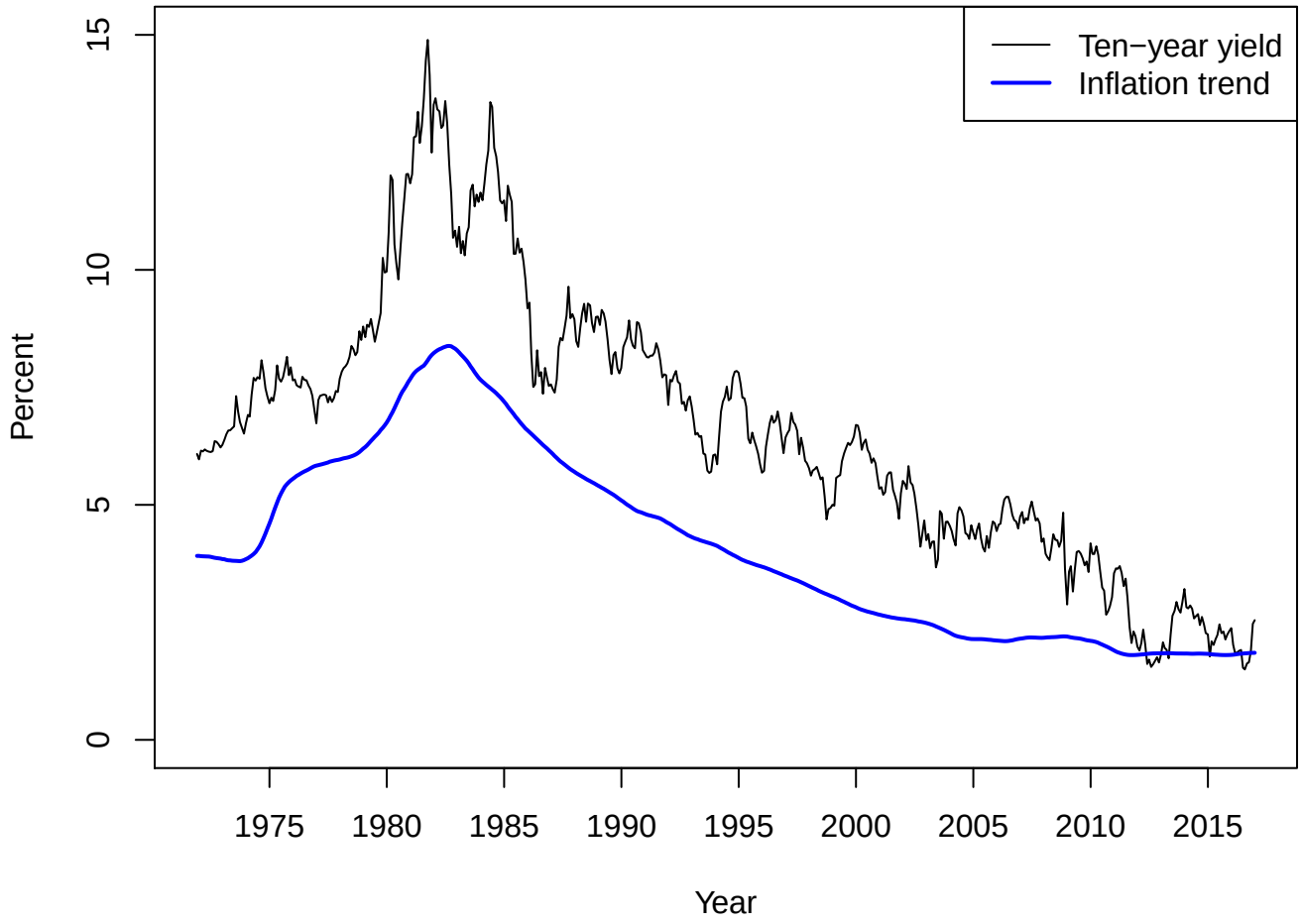
Predicting annual excess bond returns, averaged over two- through five-year bonds, using principal components (PCs) of yields. The null hypothesis is that the first three PCs contain all the relevant predictive information. The data used in the top panel is the same as in Cochrane and Piazzesi (2005)—see in particular their table 4. HAC statistics and p-values are calculated using Newey-West standard errors with 18 lags. Bootstrap distributions are obtained under the null hypothesis, using the bootstrap procedure described in the text. Under *Size* we report estimates of the size of the tests based on the bootstrap samples. Under *Power* we report power estimates using a bootstrap under the alternative hypothesis, as described in the text. p-values below 5% are emphasized with bold face.

Figure 1: Size distortions and sample size in simulation study



True size of conventional t-test of $H_0 : \beta_2 = 0$ with nominal size of 5% according to local-to-unity asymptotic distribution, for different sample sizes. DGP parameters are $\delta = 1$, $c_1 = c_2 = 0$, $\beta_0 = \beta_1 = \beta_2 = 0$, $\sigma_1 = \sigma_2 = \sigma_u = 1$, and $\rho_1 = \rho_2 = \rho$ either equal to one or 0.99. For details on the simulation study refer to text.

Figure 2: Cieslak-Povala: ten-year yield and inflation trend



Inflation trend is estimated by equation (17).

Appendix

A Derivations of theoretical results

A.1 Derivations for Section 2.2

Let $y = (y_{1+h}, y_{2+h}, ..., y_{T+h})'$ and stack x'_{1t} and x'_{2t} into $(T \times K_1)$ and $(T \times K_2)$ matrices denoted X_1 and X_2. Note that the OLS estimates of equation (1) satisfy

$$\begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} X_1'y \\ X_2'y \end{bmatrix}.$$

Premultiply the first row by $X_2'X_1(X_1'X_1)^{-1}$ and subtract the result from the second,

$$(X_2'M_1X_2)b_2 = X_2'M_1y,$$

for $M_1 = I_T - X_1(X_1'X_1)^{-1}X_1'$. Using the fact that M_1 is symmetric and idempotent we have

$$X_2'M_1X_2 = (M_1X_2)'M_1X_2 = \sum \tilde{x}_{2t}\tilde{x}'_{2t} \tag{18}$$

$$b_2 = \left(\sum \tilde{x}_{2t}\tilde{x}'_{2t}\right)^{-1}\left(\sum \tilde{x}_{2t}y_{t+h}\right). \tag{19}$$

Substituting equation (1) into (19) and using the facts that $\sum \tilde{x}_{2t}x'_{1t} = 0$ (by the orthogonality property of residuals) and that $\sum \tilde{x}_{2t}x'_{2t} = \sum \tilde{x}_{2t}\tilde{x}'_{2t}$ (again by idempotence of M_1) gives

$$b_2 = \beta_2 + \left(\sum \tilde{x}_{2t}\tilde{x}'_{2t}\right)^{-1}\left(\sum \tilde{x}_{2t}u_{t+h}\right) \tag{20}$$

from which the Wald test is

$$\begin{aligned}&(b_2 - \beta_2)'s^{-2}\textstyle\sum_{t=1}^{T}\tilde{x}_{2t}\tilde{x}'_{2t}(b_2 - \beta_2) \\ &= \left(\textstyle\sum_{t=1}^{T} u_{t+1}\tilde{x}'_{2t}\right)\left(s^2\textstyle\sum_{t=1}^{T}\tilde{x}_{2t}\tilde{x}'_{2t}\right)^{-1}\left(\textstyle\sum_{t=1}^{T}\tilde{x}_{2t}u_{t+1}\right)\end{aligned}$$

as claimed in (2)

Note that if $u|X_1, X_2 \sim N(0, \sigma_u^2 I_T)$, then K_2^{-1} times expression (2) would have an exact $F(K_2, T - K_1 - K_2)$ distribution for every sample size T and any stationary or nonstationary process for x_{2t}. Under the weaker assumption that $E(u_{t+1}|x_t, x_{t-1}, ..., x_1) = 0$ but $E(u_t|x_t, x_{t-1}, ..., x_1) \neq 0$, the Wald statistic (2) will still be asymptotically $\chi^2(K_2)$ under standard first-order stationary asymptotics, as can be seen from equation (33) below for the special case $h = 1$ and $S = \sigma_u^2 Q$. The problems arise when x_{1t} is correlated with u_t and furthermore x_t is highly persistent. In the case of unit-root processes these problems give (2) an asymptotic distribution that is not $\chi^2(K_2)$, and for near-unit-root processes they cause the small-sample distribution to be quite different from a $\chi^2(K_2)$.

The unit-root derivations this next paragraph assume a functional central limit theorem $T^{-1/2}x_{i,[T\lambda]} \Rightarrow B_i(\lambda)$ for $i = 1, 2$ with $0 \leq \lambda \leq 1$, $[T\lambda]$ the largest integer less than or equal to

$T\lambda$, $B_i(\lambda)$ K_i-dimensional Brownian motion, and $\Rightarrow$ denoting weak convergence in probability measure. From the FCLT and the Continuous Mapping Theorem,

$$\begin{aligned}\hat{A}_T &= \left[T^{-1}\int_0^1 x_{2,[T\lambda]}x'_{1,[T\lambda]}d\lambda\right]\left[T^{-1}\int_0^1 x_{1,[T\lambda]}x'_{1,[T\lambda]}d\lambda\right]^{-1} \\ &\Rightarrow \left[\int_0^1 B_2(\lambda)B_1(\lambda)'d\lambda\right]\left[\int_0^1 B_1(\lambda)B_1(\lambda)'d\lambda\right]^{-1} \\ &\equiv \tilde{A}.\end{aligned}$$

Notice that

$$\begin{aligned}\sum_{t=1}^T \tilde{x}_{2t}u_{t+1} &= \sum_{t=1}^T x_{2t}u_{t+1} - \hat{A}_T\sum_{t=1}^T x_{1t}u_{t+1} \\ &= \sum_{t=1}^T x_{2t}u_{t+1} - \sum_{t=1}^T x_{2t}x'_{1t}Z_T \qquad (21)\end{aligned}$$

for $Z_T = \left(\sum_{t=1}^T x_{1t}x'_{1t}\right)^{-1}\left(\sum_{t=1}^T x_{1t}u_{t+1}\right)$ If x_{1t} is a unit-root process that is correlated with the lag of u_{t+1}, Z_T will have a nonstandard distribution. For example, if x_{1t} is a scalar random walk with $x_{1,t+1} = x_{1t} + u_{t+1}$, then Z_T has the same distribution as $\hat{\rho}_T - 1$ where $\hat{\rho}_T$ is the OLS coefficient from a regression of $x_{1,t+1}$ on x_{1t}, a distribution with a negative bias that is well-known from unit root regressions.[45] If x_{2t} is uncorrelated with x_{1t}, then unlike the Dicky-Fuller distribution, the second term in (21) is symmetric around zero and is uncorrelated with the first term, so that the variance of $\sum_{t=1}^T \tilde{x}_{2t}u_{t+1}$ is strictly greater than that of $\sum_{t=1}^T x_{2t}u_{t+1}$.

A.2 Derivations for Section 2.3

For our local-to-unity results we assume as in Stock (1994, eq (2.17)) that $T^{-1/2}x_{i,[T\lambda]} \Rightarrow \sigma_i J_{c_i}(\lambda)$. We first note from Phillips (1988, Lemma 3.1(d)) that

$$T^{-2}\sum(x_{1t} - \bar{x}_1)^2 \Rightarrow \sigma_1^2\left\{\int_0^1 [J_{c_1}(\lambda)]^2 d\lambda - \left[\int_0^1 [J_{c_1}(\lambda)]d\lambda\right]^2\right\} = \sigma_1^2\int [J_{c_1}^\mu]^2$$

where in the sequel our notation suppresses the dependence on λ and lets $\int$ denote integration over λ from 0 to 1. The analogous operation applied to the numerator of (7) yields

$$A_T = \frac{T^{-2}\sum(x_{1t} - \bar{x}_1)(x_{2t} - \bar{x}_2)}{T^{-2}\sum(x_{1t} - \bar{x}_1)^2} \Rightarrow \frac{\sigma_1\sigma_2\int J_{c_1}^\mu J_{c_2}^\mu}{\sigma_1^2\int [J_{c_1}^\mu]^2}$$

as claimed in (7). Also

$$T^{-1/2}\bar{x}_2 = T^{-3/2}\sum x_{2t} = \int_0^1 T^{-1/2}x_{2,[T\lambda]}d\lambda \Rightarrow \sigma_2\int_0^1 J_{c_2}(\lambda)d\lambda.$$

[45] See for example Hamilton (1994, eq [17.4.7])

Since $\tilde{x}_{2t} = x_{2t} - \bar{x}_2 - A_T(x_{1t} - \bar{x}_1)$,

$$\begin{aligned} T^{-1/2}\tilde{x}_{2,[T\lambda]} &\Rightarrow \sigma_2 \left\{ J_{c_2}(\lambda) - \int_0^1 J_{c_2}(s)ds - A\left[J_{c_1}(\lambda) - \int_0^1 J_{c_1}(s)ds \right] \right\} \\ &= \sigma_2 \left\{ J_{c_2}^{\mu}(\lambda) - AJ_{c_1}^{\mu}(\lambda) \right\} = \sigma_2 K_{c_1,c_2}(\lambda) \end{aligned}$$

$$T^{-2}\sum \tilde{x}_{2t}^2 = \int_0^1 \{T^{-1/2}\tilde{x}_{2,[T\lambda]}\}^2 d\lambda \Rightarrow \sigma_2^2 \int_0^1 \{K_{c_1,c_2}(\lambda)\}^2 d\lambda. \tag{22}$$

Note we can write

$$\begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \\ u_t \end{bmatrix} = \begin{bmatrix} \sigma_1 & 0 & 0 \\ 0 & \sigma_2 & 0 \\ \delta\sigma_u & 0 & \sqrt{1-\delta^2}\sigma_u \end{bmatrix} \begin{bmatrix} v_{1t} \\ v_{2t} \\ v_{0t} \end{bmatrix}$$

where $(v_{1t}, v_{2t}, v_{0t})'$ is a martingale-difference sequence with unit variance matrix. From Lemma 3.1(e) in Phillips (1988) we see

$$\begin{aligned} T^{-1}\sum \tilde{x}_{2t}u_{t+1} &= T^{-1}\sum [x_{2t} - \bar{x}_2 - A_T(x_{1t} - \bar{x}_1)](\delta\sigma_u v_{1,t+1} + \sqrt{1-\delta^2}\sigma_u v_{0,t+1}) \\ &\Rightarrow \delta\sigma_2\sigma_u \int K_{c_1,c_2} dW_1 + \sqrt{1-\delta^2}\sigma_2\sigma_u \int K_{c_1,c_2} dW_0. \end{aligned} \tag{23}$$

Recalling (2), the t-test of a true null hypothesis about β_2 can be written as

$$\tau = \frac{\sum \tilde{x}_{2t}u_{t+1}}{\{s^2 \sum \tilde{x}_{2t}^2\}^{1/2}} = \frac{T^{-1}\sum \tilde{x}_{2t}u_{t+1}}{\{s^2 T^{-2}\sum \tilde{x}_{2t}^2\}^{1/2}} \tag{24}$$

where

$$s^2 \xrightarrow{p} \sigma_u^2. \tag{25}$$

Substituting (25), (23), and (22) into (24) produces

$$\tau \Rightarrow \frac{\sigma_2\sigma_u \left\{ \delta \int K_{c_1,c_2} dW_1 + \sqrt{1-\delta^2} \int K_{c_1,c_2} dW_0 \right\}}{\left\{ \sigma_u^2\sigma_2^2 \int (K_{c_1,c_2})^2 \right\}^{1/2}}$$

as claimed in (8).

Last we demonstrate that the variance of Z_1 exceeds unity. We can write

$$Z_1 = \frac{\int_0^1 J_{c_2}^{\mu}(\lambda) dW_1(\lambda)}{\left\{ \int_0^1 [K_{c_1,c_2}(\lambda)]^2 d\lambda \right\}^{1/2}} - \frac{A\int_0^1 J_{c_1}^{\mu}(\lambda) dW_1(\lambda)}{\left\{ \int_0^1 [K_{c_1,c_2}(\lambda)]^2 d\lambda \right\}^{1/2}} \tag{26}$$

Consider the denominator in these expressions, and note that

$$\begin{aligned}\int_0^1 [J_{c_2}^{\mu}(\lambda)]^2 d\lambda &= \int_0^1 [J_{c_2}^{\mu}(\lambda) - AJ_{c_1}^{\mu}(\lambda) + AJ_{c_1}^{\mu}(\lambda)]^2 d\lambda \\ &= \int_0^1 [K_{c_1,c_2}(\lambda)]^2 d\lambda + \int_0^1 [AJ_{c_1}^{\mu}(\lambda)]^2 d\lambda \\ &> \int_0^1 [K_{c_1,c_2}(\lambda)]^2 d\lambda\end{aligned}$$

where the cross-product term dropped out in the second equation by the definition of A in (7). This means that the following inequality holds for all realizations:

$$\left| \frac{\int_0^1 J_{c_2}^{\mu}(\lambda) dW_1(\lambda)}{\left\{ \int_0^1 [K_{c_1,c_2}(\lambda)]^2 d\lambda \right\}^{1/2}} \right| > \left| \frac{\int_0^1 J_{c_2}^{\mu}(\lambda) dW_1(\lambda)}{\left\{ \int_0^1 [J_{c_2}^{\mu}(\lambda)]^2 d\lambda \right\}^{1/2}} \right|. \tag{27}$$

Adapting the argument made in footnote 14, the magnitude inside the absolute-value operator on the right side of (27) can be seen to have a $N(0,1)$ distribution. Inequality (27) thus establishes that the first term in (26) has a variance that is greater than unity. The second term in (26) turns out to be uncorrelated with the first, and hence contributes additional variance to Z_1, although we have found that the first term appears to be the most important factor.[46] In sum, these arguments show that $\text{Var}(Z_1) > 1$.

A.3 Derivations for Section 2.4

First consider the case when $\rho_1 = \rho_2 = 1$, $\mu_1 = 0$, $\mu_2 \neq 0$, and $Corr(\varepsilon_{1t}, u_t) = 1$. Then $T^{-1/2} x_{1,[T\lambda]} \Rightarrow \sigma_1 W_1(\lambda)$ for $W_1(\lambda)$ standard Brownian motion, $T^{-1/2} \sum_{t=1}^{T} u_{t+1} \Rightarrow \sigma_1 W_1(1)$, while $x_{2t} = \mu_2 t + \sum_{s=1}^{t} \varepsilon_{2s}$ gives $T^{-1} x_{2,[T\lambda]} \Rightarrow \mu_2 \lambda$ as in Hamilton (1994, pp. 495-497). Let $x_t = (1, x_{1t}, x_{2t})'$ so $b = \beta + \left(\sum_{t=1}^{T} x_t x_t' \right)^{-1} \sum_{t=1}^{T} x_t u_{t+1}$. Define

$$\Upsilon_T = \begin{bmatrix} T^{1/2} & 0 & 0 \\ 0 & T & 0 \\ 0 & 0 & T^{3/2} \end{bmatrix}.$$

[46] These claims are based on moments of the respective functionals as estimated from discrete approximations to the Ornstein-Uhlenbeck processes.

Then very similar algebra to that in Hamilton (1994, pp. 498-500) gives

$$\begin{aligned}\Upsilon_T(b-\beta) &= \left[\Upsilon_T^{-1}\textstyle\sum x_t x_t'\Upsilon_T^{-1}\right]^{-1}\left[\Upsilon_T^{-1}\textstyle\sum x_t u_{t+1}\right] \\ &\Rightarrow \begin{bmatrix} 1 & \sigma_1\int W_1(\lambda) & \mu_2/2 \\ \sigma_1\int W_1(\lambda) & \sigma_1^2\int[W_1(\lambda)]^2 & \mu_2\sigma_1\int \lambda W_1(\lambda) \\ \mu_2/2 & \mu_2\sigma_1\int \lambda W_1(\lambda) & \mu_2^2/3 \end{bmatrix}^{-1} \begin{bmatrix} \sigma_1 W_1(1) \\ (1/2)\sigma_1^2[W^2(1)-1] \\ \mu_2\sigma_1[W_1(1)-\int W_1(\lambda)] \end{bmatrix} \\ &= \begin{bmatrix} \sigma_1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \sigma_1/\mu_2 \end{bmatrix} \begin{bmatrix} 1 & \int W_1(\lambda) & 1/2 \\ \int W_1(\lambda) & \int[W_1(\lambda)]^2 & \int \lambda W_1(\lambda) \\ 1/2 & \int \lambda W_1(\lambda) & 1/3 \end{bmatrix}^{-1} \begin{bmatrix} W_1(1) \\ (1/2)[W^2(1)-1] \\ W_1(1)-\int W_1(\lambda) \end{bmatrix}.\end{aligned}$$

Observe that the middle element, $T(b_1-\beta_1)$ is the identical distribution as that of $T(\hat{\rho}-1)$ in the Case 4 unit root distribution in Hamilton (1994, p. 499), and the t-statistic $(b_1-\beta_1)/\hat{\sigma}_{b_1}$ is identical to the Case 4 Dickey-Fuller t statistic (Hamilton (1994, eq [17.4.55])).

Consider next the case when $\rho_1 = \rho_2 = 1$, $\mu_1 \neq 0$, $\mu_2 \neq 0$, $Corr(\varepsilon_{1t}, u_t) = 1$, and $Corr(\varepsilon_{1t}, \varepsilon_{2s}) = 0$ for all s. Let's evaluate first the characteristics of a transformed regression of y_{t+1} on $\tilde{x}_t = Hx_t$ for

$$H = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & -\mu_1/\mu_2 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\tilde{b} = (\textstyle\sum \tilde{x}_t\tilde{x}_t')^{-1}\sum \tilde{x}_t y_{t+1} = (H')^{-1}b$$

$$\tilde{\beta} = (H')^{-1}\beta.$$

Then

$$\begin{aligned}\tilde{x}_{1t} &= x_{1t} - (\mu_1/\mu_2)x_{2t} \\ &= \mu_1 t + \textstyle\sum_{s=1}^t \varepsilon_{1s} - (\mu_1/\mu_2)\left(\mu_2 t + \sum_{s=1}^t \varepsilon_{2s}\right) \\ &= \textstyle\sum_{s=1}^t \varepsilon_{1s} - (\mu_1/\mu_2)\sum_{s=1}^t \varepsilon_{2s}\end{aligned}$$

and

$$\begin{aligned}T^{-1/2}\tilde{x}_{1,[T\lambda]} &\Rightarrow \sigma_1 W_1(\lambda) - (\mu_1/\mu_2)\sigma_2 W_2(\lambda) \\ &\equiv \kappa(\lambda)\end{aligned}$$

$$\Upsilon_T(\tilde{b}-\tilde{\beta}) \Rightarrow \begin{bmatrix} 1 & \int \kappa(\lambda) & \mu_2/2 \\ \int \kappa(\lambda) & \int[\kappa(\lambda)]^2 & \mu_2\int \lambda\kappa(\lambda) \\ \mu_2/2 & \mu_2\int \lambda\kappa(\lambda) & \mu_2^2/3 \end{bmatrix}^{-1} \begin{bmatrix} \sigma_1 W_1(1) \\ \sigma_1\int \kappa(\lambda)dW_1 \\ \mu_2\sigma_1[W_1(\lambda)-\int W_1(\lambda)] \end{bmatrix}.$$

The middle element, $T(\tilde{b}_1-\beta_1)$, has a distribution that approaches the Dickey-Fuller Case 4 as $\sigma_2 \to 0$ and is a related unit-root distribution for general $\sigma_2 > 0$.

Translating back in terms of the original regression, we have $b = H'\tilde{b}$, $b_1 = \tilde{b}_1$,

$$b_2 = \tilde{b}_2 - (\mu_1/\mu_2)\tilde{b}_1 = \tilde{b}_2 - (\mu_1/\mu_2)b_1$$

$$T(b_2 - \beta_2) = T(\tilde{b}_2 - \tilde{\beta}_2) - (\mu_1/\mu_2)T(b_1 - \beta_1)$$
$$\Rightarrow 0 - (\mu_1/\mu_2)T(b_1 - \beta_1)$$

since $T^{3/2}(\tilde{b}_2 - \tilde{\beta}_2) \sim O_p(1)$. Thus $b_2 - \beta_2$ has the same asymptotic distribution as $-(\mu_1/\mu_2)(b_1 - \beta_1)$, with t-tests on either b_1 or b_2 having a distribution related to the Dickey-Fuller Case 4. When x_{1t} and x_{2t} share the same trend ($\mu_1 = \mu_2$), the distribution of b_2 will simply be the negative of that of b_1.

By contrast, if we were to regress $y_{t+1} = \beta_0 + \beta_1 x_{1t} + u_{t+1}$ on x_{1t} alone, or $y_{t+1} = \beta_0 + \beta_2 x_{2t} + u_{t+1}$ on x_{2t} alone, t-tests on β_1 or β_2 would be asymptotically $N(0,1)$, from the same algebra as in Hamilton (1994, pp. 495-497). Thus for example if the true $\beta_1 \neq 0$ and $\beta_2 = 0$, when we do the regression on x_{1t} alone we would have perfectly appropriate tests about β_1, but if we add x_{2t} to the regression, tests about both β_1 and β_2 become distorted and x_{2t} could spuriously appear to be helpful in improving the estimate of β_1.

A.4 Derivations for Section 2.5

Note from (18) that

$$\sum \tilde{x}_{2t}\tilde{x}'_{2t} = \sum x_{2t}x'_{2t} - \left(\sum x_{2t}x'_{1t}\right)\left(\sum x_{1t}x'_{1t}\right)^{-1}\left(\sum x_{1t}x'_{2t}\right).$$

If x_t is covariance-stationary and ergodic for second moments,

$$\begin{aligned} T^{-1}\sum \tilde{x}_{2t}\tilde{x}'_{2t} &= T^{-1}\sum x_{2t}x'_{2t} - \left(T^{-1}\sum x_{2t}x'_{1t}\right)\left(T^{-1}\sum x_{1t}x'_{1t}\right)^{-1}\left(T^{-1}\sum x_{1t}x'_{2t}\right) \\ &\xrightarrow{p} E(x_{2t}x'_{2t}) - E(x_{2t}x'_{1t})\left[E(x_{1t}x'_{1t})\right]^{-1}E(x_{1t}x'_{2t}) \\ &= E(x_{2t}x'_{2t}) \equiv Q \end{aligned} \tag{28}$$

with the last line following from the assumption that x_{1t} and x_{2t} are uncorrelated. From (20) we also know

$$T^{1/2}(b_2 - \beta_2) = \left(T^{-1}\sum \tilde{x}_{2t}\tilde{x}'_{2t}\right)^{-1}\left(T^{-1/2}\sum \tilde{x}_{2t}u_{t+h}\right) \tag{29}$$

where

$$T^{-1/2}\sum \tilde{x}_{2t}u_{t+h} = T^{-1/2}\sum x_{2t}u_{t+h} - A_T T^{-1/2}\left(\sum x_{1t}u_{t+h}\right).$$

But if $E(x_{2t}x'_{1t}) = 0$, then $\text{plim}(A_T) = 0$, meaning

$$T^{-1/2}\sum \tilde{x}_{2t}u_{t+h} \xrightarrow{d} T^{-1/2}\sum x_{2t}u_{t+h}.$$

This will be recognized as $\sqrt{T}$ times the sample mean of a random vector with population mean zero, for which the Central Limit Theorem would take the form

$$T^{-1/2}\sum \tilde{x}_{2t}u_{t+h} \xrightarrow{d} r \sim N(0, S) \tag{30}$$

for S given in (12). Combining results (28), (29) and (30) gives (11).

To derive (13), let $b = (b'_1, b'_2)'$ denote the OLS coefficients when the regression includes both x_{1t} and x_{2t} and b_1^* the coefficients from an OLS regression that includes only x_{1t}. The

sum of squared residuals from the latter regression can be written

$$\begin{aligned} SSR_1 &= \sum (y_{t+h} - x_{1t}' b_1^*)^2 \\ &= \sum (y_{t+h} - x_t' b + x_t' b - x_{1t}' b_1^*)^2 \\ &= \sum (y_{t+h} - x_t' b)^2 + \sum (x_t' b - x_{1t}' b_1^*)^2 \end{aligned}$$

where all summations are over $t = 1, ..., T$ and the last equality follows from the orthogonality property of OLS. Thus the difference in SSR between the two regressions is

$$SSR_1 - SSR_2 = \sum (x_t' b - x_{1t}' b_1^*)^2. \tag{31}$$

It's also not hard to show that the fitted values for the full regression could be calculated as[47]

$$x_t' b = x_{1t}' b_1^* + \tilde{x}_{2t}' b_2. \tag{32}$$

Thus from (31) and (32),

$$SSR_1 - SSR_2 = \sum (\tilde{x}_{2t}' b_2)^2.$$

If the true value of β_2 is zero, then (20) becomes $b_2 = \left(\sum \tilde{x}_{2t} \tilde{x}_{2t}'\right)^{-1} \left(\sum \tilde{x}_{2t} u_{t+h}\right)$ so that

$$SSR_1 - SSR_2 = b_2' \left(\sum \tilde{x}_{2t} \tilde{x}_{2t}'\right) b_2 = \left(T^{-1/2} \sum u_{t+h} \tilde{x}_{2t}'\right) \left(T^{-1} \sum \tilde{x}_{2t} \tilde{x}_{2t}'\right)^{-1} \left(T^{-1/2} \sum \tilde{x}_{2t} u_{t+h}\right).$$

Results (28) and (30) then establish

$$SSR_1 - SSR_2 \xrightarrow{d} r' Q^{-1} r. \tag{33}$$

Recall that R^2 is defined as

$$R^2 = 1 - \frac{SSR}{\sum_{t=1}^{T} (y_{t+h} - \bar{y}_h)^2}$$

so the difference in R^2 is

$$R_2^2 - R_1^2 = \frac{(SSR_1 - SSR_2)}{\sum_{t=1}^{T} (y_{t+h} - \bar{y}_h)^2}.$$

Thus from (A.4),

$$T(R_2^2 - R_1^2) = \frac{(SSR_1 - SSR_2)}{\sum (y_{t+h} - \bar{y}_h)^2 / T} \xrightarrow{d} \frac{r' Q^{-1} r}{\gamma}$$

as claimed in (13).

[47] The easiest way to confirm the claim is to show that the residuals implied by (32) satisfy the orthogonality conditions required of the original full regression, namely, that they are orthogonal to x_{1t} and x_{2t}. That the residual $y_{t+h} - x_{1t}' b_1^* - \tilde{x}_{2t}' b_2$ is orthogonal to x_{1t} follows from the fact that $y_{t+h} - x_{1t}' b_1^*$ is orthogonal to x_{1t} by the definition of b_1^* while $\tilde{x}_{2t}$ is orthogonal to x_{1t} by the construction of $\tilde{x}_{2t}$. Likewise $y_{t+h} - \tilde{x}_{2t}' b_2$ is orthogonal to $\tilde{x}_{2t}$ by (19), and since x_{1t} is again orthogonal to $\tilde{x}_{2t}$ by the construction of $\tilde{x}_{2t}$, it follows that $y_{t+h} - x_{1t}' b_1^* - \tilde{x}_{2t} b_2$ is orthgonal to $\tilde{x}_{2t}$. Since $y_{t+h} - x_{1t}' b_1^* - \tilde{x}_{2t}' b_2$ is orthogonal to both x_{1t} and $\tilde{x}_{2t}$, it is also orthogonal to $x_{2t} = \tilde{x}_{2t} + A_T x_{1t}$.

B Alternative spanning hypotheses

Our baseline version of the spanning hypothesis is that three PCs of bond yields fully capture the information underlying expected bond returns and future interest rates, motivated by the well-known fact that three PCs capture almost all variation in the cross section of yields across maturities (Litterman and Scheinkman, 1991). But it is possible that higher-order PCs, while explaining only a miniscule share of cross-sectional variation of current yields, still contain information about expectations of future yields (see Section 6 for tests of this hypothesis). We therefore investigate two alternative versions of the spanning hypothesis, in which four or five PCs of yields span the information in the yield curve. Application of our bootstrap method to test these hypotheses is straightforward, since the only change to the approach described in Section 2.6 is that x_{1t} now contains four or five PCs. We consider these additional null hypotheses for the three empirical applications in Sections 3, 4 and 5, where macro variables are proposed as the additional predictors.

In Table B.1 we report the increase in $\bar{R}^2$ when the macro variables are added to a predictive regression of annual bond returns with $N \in \{3, 4, 5\}$ PCs of yields. We report 95%-bootstrap intervals to gauge how large of an increase in $\bar{R}^2$ would be plausible under the null hypothesis. We find that N does not affect the findings we reported in the paper: with only the exception of the original CPO sample, the increases in $\bar{R}^2$ are within the bootstrap intervals, suggesting that these increases are perfectly consistent with the spanning hypothesis.

Table B.1: Increase in $\bar{R}^2$ from addition of macro variables

	Original sample period			Later sample: 1985–2016		
	$N = 3$	$N = 4$	$N = 5$	$N = 3$	$N = 4$	$N = 5$
JPS	0.19	0.16	0.16	0.04	0.04	0.04
	(0.00, 0.23)	(0.00, 0.25)	(0.00, 0.24)	(0.00, 0.19)	(0.00, 0.20)	(0.00, 0.20)
LN	0.10	0.08	0.08	0.10	0.13	0.14
	(0.00, 0.11)	(0.00, 0.11)	(0.00, 0.12)	(0.00, 0.17)	(0.00, 0.17)	(0.00, 0.17)
CPO	0.34	0.34	0.30	0.19	0.19	0.19
	(0.00, 0.21)	(0.00, 0.21)	(0.00, 0.21)	(0.00, 0.23)	(0.00, 0.23)	(0.00, 0.23)

Increase in $\bar{R}^2$ for regressions of annual excess bond returns when macro variables are added to a specification that includes N PCs of yields. In parentheses are 95%-bootstrap intervals, obtained under the null hypothesis that the macro variables have no predictive power. The bootstrap procedure is described in the text.

In Table B.2 we consider Wald tests (for JPS and LN, using HAC standard errors) and t-tests (for CPO, using the RR approach) of the different spanning hypotheses. We report p-values of these tests using the conventional asymptotic distributions, estimates of the size of these tests based on the small-sample bootstrap distributions of the test statistics, and the bootstrap (i.e., size-corrected) p-values. The true size of the conventional five-percent tests of the spanning hypothesis is estimated to be between 32 and 52 percent. The bootstrap p-values,

which account for these enormous size distortions, are therefore much higher then conventional p-values, and often above five percent. Few noticeable differences in the bootstrap p-values arise from raising the number of yield PCs to four or five—in the original JPS sample the p-values increase and in the later LN sample they decrease. Overall, our conclusions about the robustness of published rejections of the spanning hypothesis remain unchanged when we consider versions of the spanning hypothesis with four or five PCs instead of our baseline hypothesis where three PCs span the information in the yield curve.

Table B.2: Tests of alternative spanning hypotheses

		Original sample period			Later sample: 1985–2016		
		$N=3$	$N=4$	$N=5$	$N=3$	$N=4$	$N=5$
JPS	HAC p-value	0.000	0.002	0.000	0.187	0.166	0.154
	HAC size	0.399	0.439	0.442	0.357	0.372	0.375
	BC bootstrap p-value	0.053	0.201	0.164	0.548	0.534	0.531
LN	HAC p-value	0.000	0.000	0.000	0.004	0.000	0.000
	HAC size	0.323	0.326	0.334	0.504	0.502	0.519
	Bootstrap p-value	0.008	0.002	0.005	0.279	0.040	0.045
CPO	RR p-value	0.000	0.000	0.000	0.000	0.000	0.000
	RR size	0.448	0.451	0.442	0.403	0.412	0.436
	BC bootstrap p-value	0.000	0.000	0.001	0.044	0.048	0.054

Conventional and bootstrap tests of different null hypotheses that adding macro variables to a regression with N PCs of bond yields does not increase the predictive power for annual excess bond returns. The bootstrap procedure is described in the text.

C Additional results for Joslin-Priebsch-Singleton

In Table C.1 we show additional results for the $\bar{R}^2$ in predictive regressions with three yield PCs and the macro variables GRO and INF proposed by Joslin et al. (2014). The dependent variables are the annual excess returns for bonds with maturity from two to ten years. That is, Table C.1 reports the same results for each individual bond which Table 2 reports in its top panel for the average excess return across bond maturities. To economize on space we only show the bootstrap results for the bias-corrected (BC) bootstrap.

The results in Table C.1 show that the increase in $\bar{R}^2$ when macro variables are added is often large although the spanning hypothesis is true in population. While for the two- to four-year bonds, the increase in $\bar{R}^2$ in the data is larger than the upper bound of the 95%-bootstrap interval, for the remaining bonds this statistic is within this interval, meaning that there is no statistical evidence against the spanning hypothesis.

Table C.1: Joslin-Priebsch-Singleton: $\bar{R}^2$ for excess-return regressions

		Original sample: 1985–2008			Later sample: 1985–2016		
		$\bar{R}^2_1$	$\bar{R}^2_2$	$\bar{R}^2_2 - \bar{R}^2_1$	$\bar{R}^2_1$	$\bar{R}^2_2$	$\bar{R}^2_2 - \bar{R}^2_1$
Two-year	Data	0.14	0.48	0.34	0.13	0.26	0.13
bond	BC bootstrap	0.45	0.51	0.06	0.37	0.41	0.05
		(0.10, 0.78)	(0.15, 0.80)	(0.00, 0.20)	(0.09, 0.65)	(0.14, 0.68)	(0.00, 0.17)
Three-year	Data	0.12	0.41	0.29	0.10	0.22	0.12
bond	BC bootstrap	0.39	0.45	0.06	0.31	0.36	0.05
		(0.07, 0.72)	(0.13, 0.75)	(0.00, 0.21)	(0.07, 0.59)	(0.11, 0.62)	(0.00, 0.19)
Four-year	Data	0.14	0.40	0.26	0.12	0.20	0.08
bond	BC bootstrap	0.38	0.44	0.06	0.30	0.36	0.05
		(0.08, 0.69)	(0.14, 0.72)	(0.00, 0.22)	(0.06, 0.57)	(0.11, 0.60)	(0.00, 0.19)
Five-year	Data	0.15	0.38	0.22	0.14	0.20	0.06
bond	BC bootstrap	0.35	0.41	0.06	0.28	0.33	0.06
		(0.08, 0.65)	(0.14, 0.69)	(0.00, 0.23)	(0.06, 0.54)	(0.10, 0.58)	(0.00, 0.20)
Six-year	Data	0.18	0.39	0.21	0.16	0.21	0.05
bond	BC bootstrap	0.37	0.43	0.06	0.28	0.34	0.05
		(0.10, 0.65)	(0.15, 0.69)	(0.00, 0.21)	(0.06, 0.52)	(0.10, 0.57)	(0.00, 0.19)
Seven-year	Data	0.18	0.37	0.18	0.17	0.21	0.04
bond	BC bootstrap	0.33	0.39	0.06	0.27	0.32	0.05
		(0.07, 0.59)	(0.13, 0.64)	(0.00, 0.23)	(0.05, 0.51)	(0.09, 0.55)	(0.00, 0.20)
Eight-year	Data	0.20	0.37	0.17	0.18	0.22	0.04
bond	BC bootstrap	0.33	0.39	0.06	0.26	0.32	0.05
		(0.08, 0.58)	(0.13, 0.63)	(0.00, 0.22)	(0.06, 0.50)	(0.11, 0.55)	(0.00, 0.20)
Nine-year	Data	0.22	0.39	0.16	0.19	0.23	0.03
bond	BC bootstrap	0.34	0.40	0.06	0.27	0.32	0.05
		(0.10, 0.58)	(0.15, 0.64)	(0.00, 0.22)	(0.07, 0.49)	(0.11, 0.54)	(0.00, 0.20)
Ten-year	Data	0.20	0.36	0.15	0.19	0.24	0.04
bond	BC bootstrap	0.31	0.37	0.06	0.28	0.33	0.05
		(0.07, 0.56)	(0.12, 0.61)	(0.00, 0.23)	(0.07, 0.51)	(0.11, 0.55)	(0.00, 0.19)

$\bar{R}^2$ for regressions of annual excess bond returns on three PCs of the yield curve ($\bar{R}^2_1$) and on three yield PCs together with the macro variables GRO and INF ($\bar{R}^2_2$), as well as the increase in $\bar{R}^2$. The macro data is described in the text. The results in the left half of the table are for the original sample period of Joslin et al. (2014); the data used in the right half is extended to December 2016. Each panel reports first the statistics in the data, and then the mean and the 95%-bootstrap intervals (in parentheses) of the bootstrap small-sample distribution. The bootstrap, which is explained in the text, imposes the null hypothesis that the macro variables have no predictive power.

D Additional results for Ludvigson-Ng

LN also constructed a single return-forecasting factor using a similar approach as Cochrane and Piazzesi (2005). They regressed the excess bond returns, averaged across the two- through five-year maturities, on the macro factors plus a cubed term of $F1$ which they found to be important. The fitted values of this regression produced their return-forecasting factor, denoted by $H8$. Adding $H8$ to a predictive regression that includes the Cochrane-Piazessi factor CP substantially increases the $\bar{R}^2$, and leads to a highly significant coefficient on $H8$. LN emphasized this result and interpreted it as further evidence that macro variables have predictive power beyond the information in the yield curve.

Tables D.1 and D.2 replicate LN's results for these regressions on the macro- ($H8$) and yield-based (CP) return-forecasting factors.[48] Table D.1 shows coefficient estimates and statistical significance, while Table D.2 reports $\bar{R}^2$. In LN's data, both CP and $H8$ are strongly significant with HAC p-values below 0.1%. Adding $H8$ to the regression increases the $\bar{R}^2$ by 9-11 percentage points.

One advantage of our bootstrap approach is that we can calculate the small-sample properties under the null hypothesis of complicated transformations of the original data such as these. To this end, we simply add an additional step in the construction of our artificial data by calculating CP and $H8$ in each bootstrap data set as the fitted values from preliminary regressions in the exact same way that LN did in the actual data.

Table D.1 shows that the observed increases in $\bar{R}^2$ when adding $H8$ to the regression are generally within the 95% bootstrap intervals. That is, although LN find large increases in $\bar{R}^2$ using these same regression specifications, this is not convincing evidence against the spanning hypothesis, as such increases in goodness-of-fit are perfectly plausible under the null hypothesis. And according to the bootstrap p-values for the coefficients on $H8$ in Table D.2, the macro return-forecasting factor is no longer significant at the 1% level. Furthermore, the size distortions for conventional t-tests are very substantial: a test with nominal size of five percent based on asymptotic HAC p-values has a true size of 58-61 percent. This evidence suggests that conventional HAC inference can be particularly problematic when the predictors are return-forecasting factors. Table D.2 also shows that the bootstrap test has good size and power.

We also examined the same regressions over the 1985–2016 sample period with results shown in the right half of Table D.1 and in the bottom panel of Table D.2. The observed increases in $\bar{R}^2$ are squarely in line with what we would expect under the spanning hypothesis, as indicated by the bootstrap intervals in Table D.1. The return-forecasting factors would again appear to be highly significant based on HAC p-values, but the size distortions of these tests are again very substantial and the coefficients on $H8$ are in fact not statistically significant when using the bootstrap p-values.

This evidence suggests that conventional HAC inference can be particularly problematic when the predictors are return-forecasting factors. One reason for the substantially distorted inference is their high persistence; $H8$ and CP have autocorrelations that are around 0.8, and decline only slowly with the lag length. Another reason is that the return-forecasting factors are constructed in a preliminary estimation step, which introduces additional estimation un-

[48] These results correspond to those in column 9 in tables 4-7 in LN.

Table D.1: Ludvigson-Ng: $\bar{R}^2$ for regressions with return-forecasting factors

	Original sample: 1964–2007			Later sample: 1985–2016		
	$\bar{R}^2_1$	$\bar{R}^2_2$	$\bar{R}^2_2 - \bar{R}^2_1$	$\bar{R}^2_1$	$\bar{R}^2_2$	$\bar{R}^2_2 - \bar{R}^2_1$
Two-year bond						
Data	0.31	0.42	0.11	0.16	0.22	0.06
Bootstrap	0.21	0.24	0.03	0.31	0.35	0.04
	(0.06, 0.39)	(0.09, 0.41)	(0.00, 0.11)	(0.09, 0.54)	(0.14, 0.56)	(0.00, 0.13)
Three-year bond						
Data	0.33	0.43	0.10	0.16	0.22	0.07
Bootstrap	0.20	0.24	0.04	0.29	0.33	0.04
	(0.06, 0.38)	(0.09, 0.41)	(0.00, 0.11)	(0.08, 0.51)	(0.14, 0.54)	(0.00, 0.14)
Four-year bond						
Data	0.36	0.45	0.09	0.19	0.26	0.07
Bootstrap	0.21	0.25	0.04	0.30	0.34	0.04
	(0.07, 0.39)	(0.10, 0.42)	(0.00, 0.11)	(0.10, 0.52)	(0.15, 0.54)	(0.00, 0.13)
Five-year bond						
Data	0.33	0.42	0.09	0.18	0.24	0.06
Bootstrap	0.21	0.24	0.04	0.29	0.32	0.04
	(0.06, 0.39)	(0.10, 0.41)	(0.00, 0.11)	(0.09, 0.50)	(0.14, 0.53)	(0.00, 0.14)

$\bar{R}^2$ for regressions of annual excess bond returns on yield and macro factors, as in Ludvigson and Ng (2010). $\bar{R}^2_1$ is for regressions with only the return-forecasting factor based on yield-curve information (CP), $\bar{R}^2_2$ is for regressions that also include the return-forecasting factor based on macro information ($H8$). The left side of the table shows results for the original data set used by Ludvigson and Ng (2010), and the right side shows results for a data sample that starts in 1985 and ends in 2016. We report the values of the statistics in the data, and the means and 95%-bootstrap intervals (in parentheses) for the bootstrap small-sample distributions, obtained under the null hypothesis that the macro variables have no predictive power. The bootstrap procedure is described in the text.

Table D.2: Ludvigson-Ng: statistical inference in regressions with return-forecasting factors

	Two-year bond		Three-year bond		Four-year bond		Five-year bond	
	CP	$H8$	CP	$H8$	CP	$H8$	CP	$H8$
Original sample: 1964–2007								
Coefficient	0.335	0.331	0.645	0.588	0.955	0.776	1.115	0.937
HAC t-statistic	4.429	4.331	4.666	4.491	4.765	4.472	4.371	4.541
HAC p-value	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
Bootstrap 5% c.v.		3.857		3.968		3.965		3.998
Bootstrap p-value		**0.019**		**0.021**		**0.023**		**0.019**
Size								
HAC		0.579		0.612		0.610		0.594
Bootstrap		0.049		0.059		0.054		0.049
Power								
Bootstrap		0.621		0.573		0.555		0.521
Later sample: 1985–2016								
Coefficient	0.363	0.333	0.678	0.663	1.101	0.934	1.314	1.146
HAC statistic	2.746	2.768	2.556	3.073	2.933	3.308	2.837	3.379
HAC p-value	**0.006**	**0.006**	**0.011**	**0.002**	**0.004**	**0.001**	**0.005**	**0.001**
Bootstrap 5% c.v.		4.182		4.172		4.158		4.160
Bootstrap p-value		0.271		0.199		0.153		0.134

Predictive regressions for annual excess bond returns, using return-forecasting factors based on yield-curve information (CP) and macro information ($H8$), as in Ludvigson and Ng (2010). The first panel shows the results for their original data and sample period; the second panel uses a data sample that starts in 1985 and ends in 2016. HAC t-statistics and p-values are calculated using Newey-West standard errors with 18 lags. We obtain bootstrap small-sample distributions of the t-statistics under the null hypothesis that macro factors and hence $H8$ have no predictive power, and report the bootstrap critical values (c.v.'s) and p-values, as well as estimates of the true size of conventional HAC t-tests and the bootstrap tests with 5% nominal coverage (see notes to Table 3). We also report estimates of the power of the bootstrap tests. The bootstrap procedure is described in the text. p-values below 5% are emphasized with bold face.

certainty not accounted for by conventional inference. We recommend that researchers use our bootstrap in such a setting to accurately carry out inference. Here we conclude that LN's macro return-forecasting factor exhibits only very tenuous predictive power, much weaker than indicated by LN's original analysis, which disappears completely over a different sample period.

E Bond supply: Greenwood-Vayanos

A large literature studies the effects of the supply of bonds on prices and yields, including the recent contributions of Hamilton and Wu (2012) and Greenwood and Vayanos (2014). Theoretical and empirical work has demonstrated that bond supply is related to bond yields and returns. But do measures of Treasury bond supply contain predictive power for bond returns that is not already reflected in the yield curve? The answer appears to be yes: Greenwood and Vayanos (2014) (henceforth GV) found that variation in their measure of bond supply, a maturity-weighted debt-to-GDP ratio, predicts yields and bond returns, and that this holds true even controlling for yield curve information such as the term spread. Here we investigate whether this result is robust and holds up to closer scrutiny. The sample period used in GV is 1952 to 2008.

We are most interested in those regression specifications estimated by GV that control for the information in the yield curve. We first reproduce, in the top panel of Table E.1, their baseline specification in which the one-year return on a long-term bond is predicted using the one-year yield and bond supply measure alone. The second panel includes the spread between the long-term and one-year yield as an additional explanatory variable.[49] Like GV we use Newey-West standard errors with 36 lags. If we interpreted the HAC t-test using the conventional asymptotic critical values, the coefficient on bond supply is significant in the baseline regression in the top panel. When the yield spread is included in the regression, this coefficient is marginally insignificant, with a p-value of 5.8%.

The bond return that GV used as the dependent variable in these regressions is for a hypothetical long-term bond with a 20-year maturity. We cannot apply our bootstrap procedure here because this bond return is not constructed from the observed yield curve.[50]

We consider two additional regression specifications that are relevant in this context. The first specification controls for information in the yield curve by including, instead of a single term spread, the first three PCs of observed yields.[51] It also subtracts the one-year yield from the bond return in order to yield an excess return. Both of these changes make this specification more closely comparable to those in the literature. The results are reported in the third panel of Table E.1. Again, the coefficient on bond supply is only marginally significant for the HAC t-test.

Finally, we consider a specification where the one-year excess return, averaged across two- though five-year maturities, is regressed on yield PCs and the measure of bond supply. The last panel of Table E.1 shows that in this case, the coefficient on bond supply is insignificant according to the conventional Newey-West t-test. In this last regression, which includes PCs

[49] These estimates are in GV's table 5, rows 1 and 6. Their baseline results are also in their table 2.

[50] GV obtained this series from Ibbotson Associates.

[51] These PCs are calculated from the observed Fama-Bliss yields with one- through five-year maturities.

Table E.1: Greenwood-Vayanos: predictive power of Treasury bond supply

	One-year yield	Term spread	$PC1$	$PC2$	$PC3$	Bond supply
Dependent variable: return on long-term bond						
Coefficient	1.212					0.026
HAC t-statistic	2.853					3.104
HAC p-value	**0.004**					**0.002**
Dependent variable: return on long-term bond						
Coefficient	1.800	2.872				0.014
HAC t-statistic	5.208	4.596				1.898
HAC p-value	**0.000**	**0.000**				0.058
Dependent variable: excess return on long-term bond						
Coefficient			0.168	5.842	-6.089	0.013
HAC t-statistic			1.457	4.853	1.303	1.862
HAC p-value			0.146	**0.000**	0.193	0.063
Dependent variable: avg. excess return for 2-5 year bonds						
Coefficient			0.085	1.669	-4.632	0.004
HAC statistic			1.270	3.156	2.067	1.154
HAC p-value			0.204	**0.002**	**0.039**	0.249
Bootstrap 5% c.v.						3.199
Bootstrap p-value						0.468

Predictive regressions for annual bond returns using Treasury bond supply, as in Greenwood and Vayanos (2014) (GV). The coefficients on bond supply in the first two panels are identical to those reported in rows (1) and (6) of Table 5 in GV. HAC t-statistics and p-values are constructed using Newey-West standard errors with 36 lags, as in GV. The last panel includes bootstrap critical values and p-values using small-sample distributions generated under the null hypothesis that bond supply does not contain additional predictive power—the bootstrap procedure is described in the text. The last two rows in each panel report p-values for t-tests using the methodology of Ibragimov and Müller (2010), splitting the sample into either 8 or 16 blocks. The sample period is 1952 to 2008. p-values below 5% are emphasized with bold face.

Table F.1: Cooper-Priestley: predictive power of the output gap

	gap	$\tilde{CP}$	CP	$PC1$	$PC2$	$PC3$
Coefficient	-0.126					
OLS t-statistic	3.225					
HAC t-statistic	1.077					
HAC p-value	0.282					
Coefficient	-0.120	1.588				
OLS t-statistic	3.479	13.541				
HAC t-statistic	1.244	4.925				
HAC p-value	0.214	**0.000**				
Coefficient	0.113		1.612			
OLS t-statistic	2.940		13.831			
HAC t-statistic	1.099		5.059			
HAC p-value	0.273		**0.000**			
Coefficient	0.147			0.001	0.043	-0.067
OLS t-statistic	3.524			4.359	11.506	3.690
HAC t-statistic	1.306			1.332	4.363	2.508
HAC p-value	0.192			0.183	**0.000**	0.012
Bootstrap 5% c.v.	2.843					
Bootstrap p-value	0.354					

Predictive regressions for the one-year excess return on a five-year bond using the output gap, as in Cooper and Priestley (2008) (CPR). $\tilde{CP}$ is the Cochrane-Piazzesi factor after orthogonalizing it with respect to gap, whereas CP is the usual Cochrane-Piazzesi factor. For the predictive regression, gap is lagged one month, as in CPR. HAC standard errors are based on the Newey-West estimator with 22 lags. The bootstrap procedure, which does not include bias correction, is described in the main text. The sample period is 1952 to 2003. p-values below 5% are emphasized with bold face.

and a conventional excess bond return, we can also use our bootstrap procedure. We find that the bootstrap p-value is substantially higher than the conventional p-value. The bond supply variable has a first-order autocorrelation is 0.998, which causes substantial size distortions for the conventional t-test in this and in the other regression specifications.

Overall, we find that the results in GV do not constitute evidence against the spanning hypothesis. While bond supply exhibits a strong empirical link with interest rates, its predictive power for future yields and returns seems to be fully captured by the current yield curve.

F Output gap: Cooper-Priestley

Another widely cited study that appears to provide evidence of predictive power of macro variables for asset prices is Cooper and Priestley (2008) (henceforth CPR). This paper focuses on one particular macro variable as a predictor of stock and bond returns, namely the output gap, which is a key indicator of the economic business cycle. The authors concluded that "the output gap can predict next year's excess returns on U.S. government bonds" (p. 2803). Furthermore, they also claimed that some of this predictive power is independent of the information in the yield curve, and implicitly rejected the spanning hypothesis (p. 2828).

Like CPR we use $x_{2t} = gap_{t-1}$, the output gap at date $t-1$, measured as the deviation of the Fed's Industrial Production series from a quadratic time trend.[52] CPR lagged their measure by one month to account for the publication lag of the Fed's Industrial Production data. Table F.1 shows our results for predictions of the excess return on the five-year bond; the results for other maturities closely parallel these. The top two panels correspond to the regression specifications that CPR estimated.[53] In the first specification, the only predictor is gap_{t-1}. The second specification also includes $\tilde{CP}_t$, which is the Cochrane-Piazzesi factor CP_t after it is orthogonalized with respect to gap_t.[54] We obtain coefficients and $\bar{R}^2$ that are close to those published in CPR. We calculate both OLS and HAC t-statistics, where in the latter case we use Newey-West with 22 lags as described by CPR. Our OLS t-statistics are very close to the published numbers, and according to these the coefficient on gap_{t-1} is highly significant. It appears that CPR may have mistakenly reported the OLS instead of the Newey-West t-statistics, which is about a third as large as the OLS t-statistics and implies that the coefficient on gap is far from significant, with p-values above 20%.

Importantly, neither of the specifications in CPR can be used to test the spanning hypothesis, because the CP factor is first orthogonalized with respect to the output gap. This defeats the purpose of controlling for yield-curve information, since any predictive power that is shared by the CP factor and gap will be exclusively attributed to the latter. In particular, finding a significant coefficient on gap in a regression with $\tilde{CP}$ cannot justify the conclusion that "gap is capturing risk that is independent of the financial market-based variable CP" (p. 2828). One way to test the spanning hypothesis is to include CP instead of $\tilde{CP}$, and we report these results in the third panel of Table F.1. In this case, the coefficient on gap switches to a positive sign, and its Newey-West t-statistic remains insignificant.

Our preferred specification includes the first three PCs of the yield curve—see the last panel of Table F.1. The predictor gap is highly persistent, with a first-order autocorrelation coefficient of 0.975, so there are likely small-sample inference problems. Hence we also include results for robust inference using the bootstraptest. The gap variable has a positive coefficient with a HAC p-value of 19%, which rises to 36% when using our bootstrap procedure. The conventional HAC t-test is substantially oversized, as evident by the bootstrap critical value that substantially exceeds the conventional critical value. Overall, we do not find any evidence that the output gap predicts excess bond returns.

[52] We thank Richard Priestley for sending us this real-time measure of the output gap.

[53] The relevant results in CPR are in the top panel of their table 9.

[54] Note that the predictors $\tilde{CP}_t$ and gap_{t-1} are therefore not completely orthogonal.