Robust Bond Risk Premia*

Michael D. Bauer† and James D. Hamilton‡

April 16, 2015
Revised: January 20, 2016

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 evidence in five published studies, find most rejections of the spanning hypothesis to be spurious, and conclude that the current consensus is wrong. Only the level and the slope of the yield curve are robust predictors of bond returns.

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 John Cochrane, Graham Elliott, Robin Greenwood, Helmut Lütkepohl, Ulrich Müller, Hashem Pesaran and Glenn Rudebusch for useful suggestions, conference participants at the 7th Annual Volatility Institute Conference at the NYU Stern School of Business and at the NBER 2015 Summer Institute, as well as seminar participants at the Federal Reserve Bank of Boston and the Free University of Berlin for helpful comments, and Javier Quintero and Simon Riddell for excellent research assistance.

†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

The nominal yield on a 10-year U.S. Treasury bond has been below 2% much of the time since 2011, a level never seen previously. To what extent does this represent unprecedently low expected interest rates extending through the next decade, and to what extent does it reflect an unusually low risk premium resulting from a flight to safety and large-scale asset purchases by central banks that depressed the long-term yield? Finding the answer is a critical input for monetary policy, investment strategy, and understanding the lasting consequences of the financial and economic disruptions of 2008.

In principle one can measure the risk premium by the difference between the current long rate and the expected value of future short rates. But what information should go into constructing that expectation of future short rates? A powerful argument can be made 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 short-term interest rates and determine bond risk premia. Hence current yields are necessarily a function of z_t, reflecting the general fact that current asset prices incorporate all current information. This suggests that we may be able to back out the state vector z_t from the observed yield curve.[1] The "invertibility" or "spanning" hypothesis states that the current yield curve contains all the information that is useful for predicting future interest rates or determining risk premia. Notably, under this hypothesis, the yield curve is first-order Markov.

It has long been recognized that three yield-curve factors, such as the first three principal components (PCs) of yields, can provide an excellent summary of the information in the entire yield curve (Litterman and Scheinkman, 1991). While it is clear that these factors, which are commonly labeled level, slope, and curvature, explain almost all of the cross-sectional variance of yields, it is less clear whether they completely capture the relevant information for forecasting future yields and estimating bond risk premia. In this paper we investigate what we will refer to as the "spanning hypothesis" which holds that all the relevant information for predicting future yields and returns is spanned by the level, slope and curvature of the yield curve. This hypothesis differs from the claim that the yield curve follows a first-order Markov process, as it adds the assumption that only these three yield-curve factors are useful in forecasting. For example, if higher-order yield-curve factors such as the 4th and 5th PC are informative about predicting yields and returns, yields would still be Markov, but the spanning hypothesis, as we define it here, would be violated. Note also that the spanning

[1] Specifically, this invertibility requires that (a) we observe at least as many yields as there are state variables in z_t, and (b) there are no knife-edge cancellations or pronounced nonlinearities; see for example Duffee (2013b).

hypothesis is much less restrictive than the expectations hypothesis, which states that bond risk premia are constant and hence excess bond returns are not predictable.

The question whether the spanning hypothesis is valid is of crucial importance for finance and macroeconomics. If it is valid, then the estimation of monetary policy expectations and bond risk premia would not require any data or models involving macroeconomic series, other asset prices or quantities, volatilities, or survey expectations. Instead, all the necessary information is in the shape of the current yield curve, summarized by the level, slope, and curvature. If, however, the spanning hypothesis is violated, then this would seem to invalidate a large body of theoretical work in asset pricing and macro-finance, since models in this literature generally imply that the state variables are spanned by the information in the term structure of interest rates.[2] A growing literature on yield curve modeling is based on the premise that it is undesirable and potentially counterfactual to assume spanning.[3] There appears to be a consensus, reflected in recent review articles by Gürkaynak and Wright (2012) and Duffee (2013a), that the spanning question is a central issue in macro-finance.

A number of recent studies have produced evidence that appears to contradict the spanning hypothesis. 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), higher-order (fourth and fifth) PCs of bond yields (Cochrane and Piazzesi, 2005), the output gap (Cooper and Priestley, 2008), and measures of Treasury bond supply (Greenwood and Vayanos, 2014). The results in each of these studies suggest that there might be unspanned or hidden information that is not captured by the level, slope, and curvature of the current yield curve but that is useful for forecasting.

But the predictive regressions underlying all these results have a number of problematic features. First, the predictive variables are typically very persistent, in particular in relation to the small available sample sizes. Second, some of these predictors summarize the information in the current yield curve, and therefore are generally correlated with lagged forecast errors, i.e., they violate the condition of strict exogeneity. In such a setting, tests of the spanning hypothesis are necessarily oversized in small samples, as we show both analytically and using simulations. Third, the dependent variable is typically a bond return over an annual holding period, which introduces substantial serial correlation in the prediction errors. This worsens the size distortions and leads to large increases in R^2 even if irrelevant predictors

[2]Key contributions to this large literature include Wachter (2006), Piazzesi and Schneider (2007), Rudebusch and Wu (2008), and Bansal and Shaliastovich (2013). For a recent example see Swanson (2015).

[3]Examples are Wright (2011), Chernov and Mueller (2012), Priebsch (2014), and Coroneo et al. (2015).

are included.[4] We demonstrate that the procedures commonly used for inference about the spanning hypothesis do not appropriately address these issues, and are subject to serious small-sample distortions.

We propose two procedures that give substantially more robust small-sample inference. The first is 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.[5] 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. Our procedure notably differs from the bootstrap approach often employed in this literature, which generates artificial data under the expectations hypothesis.[6] This reveals that the conventional tests reject the true null much too often. We show for example that the tests employed by Ludvigson and Ng (2009), which are intended to have a nominal size of five percent, can have a true size of up to 54%. We then ask whether under the null it would be possible to observe similar patterns of predictability as researchers have found in the data. We find that this is indeed the case, meaning that much of the above-cited evidence against the spanning hypothesis is in fact spurious. These results provide a strong caution against using conventional tests, and we recommend that researchers instead use the bootstrap procedure proposed in this paper. Despite the usual technical concerns about bootstrapping near-nonstationary variables, we present evidence that this procedure performs well in small small samples.

A second procedure that we propose for inference in this context is the approach for robust testing of Ibragimov and Müller (2010). The approach is to split the sample into subsamples, to estimate coefficients separately in each of these, and then to perform a simple t-test on the coefficients across subsamples. We have found this approach to have excellent size and power properties in settings similar to the ones encountered by researchers testing for predictive power for interest rates and bond returns. Applying this type of test to the predictive regressions

[4] Lewellen et al. (2010) demonstrated that high R^2 in cross-sectional return regressions are, for different reasons, often unconvincing evidence of true predictability.

[5] We consider bias-corrected estimation of the VAR, in light of the high persistence of the PCs.

[6] This approach has been used, for example, by Bekaert et al. (1997), Cochrane and Piazzesi (2005), Ludvigson and Ng (2009, 2010), and Greenwood and Vayanos (2014).

for excess bond returns studied in the literature, we find that the only robust predictors are the level and the slope of the yield curve.

After revisiting the evidence in the five influential papers cited above we draw two conclusions. First, the claims going back to Fama and Bliss (1987) and Campbell and Shiller (1991) that excess returns can be predicted from the level and slope of the yield curve remain quite robust. Second, the newer evidence on the predictive power of macro variables, higher-order PCs of the yield curve, or other variables, is subject to more serious econometric problems and appears weaker and much less robust.We further demonstrate that this predictive power is substantially weaker in samples that include subsequent data than in the samples originally analyzed. Overall, we do not find convincing evidence to reject the hypothesis that the current yield curve, and in particular three factors summarizing this yield curve, contains all the information necessary to infer interest rate forecasts and bond risk premia. In other words, the spanning hypothesis cannot be rejected, and the Markov property of the yield curve seems alive and well.

Our paper is related mainly to two strands of literature. The first is the literature on the spanning hypothesis, and most relevant studies were cited above. Bauer and Rudebusch (2015) also question the evidence against spanning, by showing that conventional macro-finance models can generate data in which the spanning hypothesis is spuriously rejected. Our paper is also related to the econometric literature on spurious results in return regressions. Mankiw and Shapiro (1986), Cavanagh et al. (1995), Stambaugh (1999) and Campbell and Yogo (2006), among others, studied short-horizon return predictability with a regressor that is not strictly exogenous. We point out a related econometric issue in bond return regressions, which is however distinct from Stambaugh bias. Ferson et al. (2003) and Deng (2013) studied the size distortions in a setting that is different from ours and more relevant for stock returns, namely when returns have an unobserved persistent component. In contrast to these studies, we focus on the econometric problems that arise in tests of the spanning hypothesis. In addition, we propose simple, easily implementable solutions to these problems.

2 Inference about the spanning hypothesis

The evidence against the spanning hypothesis in all of the studies cited in the introduction 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) that we wish to predict, 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.[7] 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.

The evidence produced in these studies comes in two forms, the first based on simple descriptive statistics such as how much the R^2 of the regression increases when the variables x_{2t} are added and the second from formal statistical tests of the hypothesis that $\beta_2 = 0$. In this section we show how key features of the specification can matter significantly for both forms of evidence. In Section 2.1 we show how serial correlation in the error term u_t and the proposed predictors x_{2t} can give rise to a large increase in R^2 when x_{2t} is added to the regression even if it is no help in predicting y_{t+h}. In Section 2.2 we show the consequences of lack of strict exogeneity of x_{1t}, which is necessarily correlated with u_t since it contains information in current yields. When x_{1t} and x_{2t} are highly persistent processes, as is usually the case in practice, conventional tests of H_0 generally will exhibit significant size distortions in finite samples. We then propose methods for robust inference about bond return predictability in Sections 2.3 and 2.4.

2.1 Implications of serially correlated errors based on first-order asymptotics

Our first observation is that in regressions in which x_{1t} and x_{2t} are strongly persistent and the error term is serially correlated—as is always the case with overlapping bond returns—we should not be surprised to see substantial increases in R^2 when x_{2t} is added to the regression even if the true coefficient is zero. It is well known that in small samples serial correlation in the residuals can increase both the bias as well as the variance of a regression R^2 (see for example Koerts and Abrahamse (1969) and Carrodus and Giles (1992)). To see how much

[7] We will always sign the PCs so that the yield with the longest maturity loads positively on all PCs. As a result, PC1 and PC2 correspond to what are commonly referred to as "level" and "slope" of the yield curve.

difference this could make in the current setting, consider the unadjusted R^2 defined as

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

where SSR denotes the regression sum of squared residuals. The increase in R^2 when x_{2t} is added to the regression is thus given by

$$R_2^2 - R_1^2 = \frac{(SSR_1 - SSR_2)}{\sum_{t=1}^{T}(y_{t+h} - \bar{y}_h)^2}. \tag{3}$$

We show in Appendix A that when x_{1t}, x_{2t}, and u_{t+h} are stationary and satisfy standard regularity conditions, if the null hypothesis is true ($\beta_2 = 0$) and the extraneous regressors are uncorrelated with the valid predictors ($E(x_{2t}x_{1t}') = 0$), then

$$T(R_2^2 - R_1^2) \xrightarrow{d} r'Q^{-1}r/\gamma \tag{4}$$

$$\gamma = E[y_t - E(y_t)]^2$$

$$r \sim N(0, S), \tag{5}$$

$$Q = E(x_{2t}x_{2t}') \tag{6}$$

$$S = \sum_{v=-\infty}^{\infty} E(u_{t+h}u_{t+h-v}x_{2t}x_{2,t-v}'). \tag{7}$$

Result (4) implies that the difference $R_2^2 - R_1^2$ itself converges in probability to zero under the null hypothesis that x_{2t} does not belong in the regression, meaning that the two regressions asymptotically should have the same R^2.

In a given finite sample, however, R_2^2 is larger than R_1^2 by construction, and the above results give us an indication of how much larger it would be in a given finite sample. If $x_{2t}u_{t+h}$ is serially uncorrelated, then (7) simplifies to $S_0 = E(u_{t+h}^2 x_{2t}x_{2t}')$. On the other hand, 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.

When the dependent variable y_{t+h} is a multi-period bond return, then the error term is necessarily serially correlated. In our empirical applications, y_{t+h} will typically be the h-period excess return on an n-period bond,

$$y_{t+h} = p_{n-h,t+h} - p_{nt} - hi_{ht}, \tag{8}$$

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 $v = 0, \ldots, h-1$, due to the overlapping observations. At the same time, the explanatory variables x_{2t} often are highly serially correlated, so $E(x_{2t}x'_{2,t-v}) \neq 0$. Thus even if x_{2t} is completely independent of u_t at all leads and lags, the product will be highly serially correlated,

$$E(u_{t+h}u_{t+h-v}x_{2t}x'_{2,t-v}) = E(u_t u_{t-v})E(x_{2t}x'_{2,t-v}) \neq 0.$$

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.[8]

These results on the asymptotic distribution of $R_2^2 - R_1^2$ could be used to design a test of H_0. However, we show in the next subsection that in small samples additional problems arise from the persistence of the predictors, with the consequence that the bias and variability of $R_2^2 - R_1^2$ can be even greater than predicted by (4). For this reason, in this paper we will rely on bootstrap approximations to the small-sample distribution of the statistic $R_2^2 - R_1^2$, and demonstrate that the dramatic values sometimes reported in the literature are not implausible under the spanning hypothesis.

Serial correlation of the residuals also affects the sampling distribution of the OLS estimate of β_2. In Appendix A we verify using standard algebra that under the null hypothesis $\beta_2 = 0$ the OLS estimate b_2 can be written as

$$b_2 = \left(\sum_{t=1}^{T} \tilde{x}_{2t}\tilde{x}'_{2t}\right)^{-1} \left(\sum_{t=1}^{T} \tilde{x}_{2t}u_{t+h}\right) \tag{9}$$

where $\tilde{x}_{2t}$ denotes the sample residuals from OLS regressions of x_{2t} on x_{1t}:

$$\tilde{x}_{2t} = x_{2t} - A_T x_{1t} \tag{10}$$

[8]The same conclusions necessarily also hold for the adjusted $\bar{R}^2$ defined as

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

for k_i the number of coefficients estimated in model i, from which we see that

$$T(\bar{R}_2^2 - \bar{R}_1^2) = \frac{[T/(T-k_1)]SSR_1 - [T/(T-k_2)]SSR_2}{\sum_{t=1}^{T}(y_{t+h} - \bar{y}_h)^2/(T-1)}$$

which has the same asymptotic distribution as (4). In our small-sample investigations below, we will analyze either R^2 or $\bar{R}^2$ as was used in the original study that we revisit.

$$A_T = \left(\sum_{t=1}^{T} x_{2t} x_{1t}'\right) \left(\sum_{t=1}^{T} x_{1t} x_{1t}'\right)^{-1}. \tag{11}$$

If x_{2t} and x_{1t} are stationary and uncorrelated with each other, as the sample size grows, $A_T \xrightarrow{p} 0$ and b_2 has the same asymptotic distribution as

$$b_2^* = \left(\sum_{t=1}^{T} x_{2t} x_{2t}'\right)^{-1} \left(\sum_{t=1}^{T} x_{2t} u_{t+h}\right), \tag{12}$$

namely

$$\sqrt{T} b_2 \xrightarrow{d} N(0, Q^{-1} S Q^{-1}). \tag{13}$$

with Q and S the matrices defined in (6) and (7). Again we see that positive serial correlation causes S to exceed the value S_0 that would be appropriate for serially uncorrelated residuals. In other words, serial correlation in the error term increases the sampling variability of the OLS estimate b_2.

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, in practice different HAC estimators of S can lead to substantially different empirical conclusions (Müller, 2014). Moreover, we show in the next subsection that even if the population value of S were known with certainty, expression (13) can give a poor indication of the true small-sample variance. We further demonstrate empirically in the subsequent sections that this is a serious problem when carrying out inference about bond return predictability.

2.2 Small-sample implications of lack of strict exogeneity

A second feature of the studies examined in this paper is that the valid explanatory variables x_{1t} are correlated with lagged values of the error term. That is, they are only weakly but not strictly exogenous. In addition, x_{1t} and x_{2t} are highly serially correlated. We will show that this can lead to substantial size distortions in tests of $\beta_2 = 0$. The intuition of our result is the following: As noted above, the OLS estimate of β_2 in (1), b_2, can be thought of as being implemented in three steps: (i) regress x_{2t} on x_{1t}, (ii) regress y_{t+h} on x_{1t}, and (iii) regress the residuals from (ii) on the residuals of (i). When x_{1t} and x_{2t} are highly persistent, the auxiliary regression (i) behaves like a spurious regression in small samples, causing $\sum \tilde{x}_{2t} \tilde{x}_{2t}'$ in (9) to be significantly smaller than $\sum x_{2t} x_{2t}'$ in (12). When there is correlation between x_{1t} and u_t, this causes the usual asymptotic distribution to underestimate significantly the true variability of b_2. As a consequence, the t-test for $\beta_2 = 0$ rejects the true null too often. In the following, we demonstrate exactly why this occurs, first theoretically using local-to-unity asymptotics,

and then in small-sample simulations.

The issue we raise has to our knowledge not previously been recognized. Mankiw and Shapiro (1986) and Stambaugh (1999) studied tests of the hypothesis $\beta_1 = 0$ in a regression of y_{t+1} on x_{1t}, where the regressors x_{1t} are not strictly exogenous, and documented that when x_{1t} is persistent this leads to small-sample coefficient bias in the OLS estimate of β_1.[9] By contrast, in our setting there is no coefficient bias present in estimates of β_2, and it is instead the inaccuracy of the standard errors, which we will refer to as "standard error bias," that distorts the results of conventional inference. Another related line of work is by Ferson et al. (2003) and Deng (2013), who studied predictions of returns that have a persistent component that is unobserved. In our notation, their setting corresponds to the case where both x_{1t} and x_{2t} are strictly exogenous, x_{1t} is unobserved, and returns are predicted using x_{2t}. For predictive regressions of bond returns, however, we do have estimates of the persistent return component based on information in the current yield curve, x_{1t}, and instead the resulting lack of strict exogeneity causes a separate econometric problem from that considered by Ferson et al. (2003) and Deng (2013).

2.2.1 Theoretical anlysis using local-to-unity asymptotics

We now demonstrate where the problem arises in the simplest example of our setting. Suppose that x_{1t} and x_{2t} are scalars that follow independent highly persistent processes,

$$x_{i,t+1} = \rho_i x_{it} + \varepsilon_{i,t+1} \quad i = 1, 2 \tag{14}$$

where ρ_i is close to one. Consider the consequences of OLS estimation of (1) in the special case where $h = 1$:

$$y_{t+1} = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + u_{t+1}. \tag{15}$$

We assume that $(\varepsilon_{1t}, \varepsilon_{2t}, u_t)'$ follows a martingale difference sequence with finite fourth moments and variance matrix

$$V = E \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \\ u_t \end{bmatrix} \begin{bmatrix} \varepsilon_{1t} & \varepsilon_{2t} & u_t \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & 0 & \delta\sigma_1\sigma_u \\ 0 & \sigma_2^2 & 0 \\ \delta\sigma_1\sigma_u & 0 & \sigma_u^2 \end{bmatrix}. \tag{16}$$

Thus x_{1t} is not strictly exogenous when the correlation δ is nonzero. Note that for any δ, $x_{2t}u_{t+1}$ is serially uncorrelated and the standard OLS t-test of $\beta_2 = 0$ asymptotically has a $N(0, 1)$

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

distribution when using the conventional first-order asymptotic approximation. This simple example illustrates the problems in a range of possible settings for yield-curve forecasting. In particular, if $\text{Var}(u_{t+1})$ substantially exceeds $\text{Var}(\beta_1 x_{1t})$, y_t could be viewed as a (one-period) bond return, where $\beta_1 x_{1t}$ is a persistent component of the return that is small relative to the size of y_{t+1}.

One device for seeing how the results in a finite sample of some particular size T likely 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):

$$x_{i,t+1} = (1 + c_i/T)x_{it} + \varepsilon_{i,t+1} \quad i = 1, 2. \tag{17}$$

For example, if our data come from a sample of size $T = 100$ when $\rho_i = 0.95$, the idea is to represent this with a value of $c_i = -5$ in (17). The claim is that analyzing the properties as $T \to \infty$ of a model characterized by (17) with $c_i = -5$ gives a better approximation to the actual distribution of regression statistics in a sample of size $T = 100$ and $\rho_i = 0.95$ than is provided by the first-order asymptotics used in the previous subsection which treat ρ_i as a constant when $T \to \infty$; see for example Chan (1988) and Nabeya and Sørensen (1994). The local-to-unity 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

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

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

with $W_1(\lambda)$ and $W_2(\lambda)$ denoting independent standard Brownian motion. When $c_i = 0$, (17) 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: $J_0(\lambda) = W(\lambda)$.

Applying local-to-unity asymptotics to our setting reveals the basic econometric problem. We show in Appendix B that under local-to-unity asymptotics the coefficient from a regression of x_{2t} on x_{1t} has the following limiting distribution:

$$A_T = \frac{\sum(x_{1t} - \bar{x}_1)(x_{2t} - \bar{x}_2)}{\sum(x_{1t} - \bar{x}_1)^2} \Rightarrow \frac{\sigma_2 \int_0^1 J_{c_1}^{\mu}(\lambda) J_{c_2}^{\mu}(\lambda) d\lambda}{\sigma_1 \int_0^1 [J_{c_1}^{\mu}(\lambda)]^2 d\lambda} \equiv (\sigma_2/\sigma_1) A, \tag{18}$$

where we have defined A to be the random variable in the middle expression. 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 proportional to A. Consequently, the t-statistic for $\beta_2 = 0$ can have a very different distribution from that predicted using first-order asymptotics. We demonstrate in Appendix B that this t-statistic has a local-to-unity asymptotic distribution under the null hypothesis that is given by

$$\frac{b_2}{\{s^2/\sum \tilde{x}_{2t}^2\}^{1/2}} \Rightarrow \delta Z_1 + \sqrt{1-\delta^2} Z_0 \tag{19}$$

$$Z_1 = \frac{\int_0^1 K_{c_1,c_2}(\lambda) dW_1(\lambda)}{\left\{\int_0^1 [K_{c_1,c_2}(\lambda)]^2 d\lambda\right\}^{1/2}} \tag{20}$$

$$Z_0 = \frac{\int_0^1 K_{c_1,c_2}(\lambda) dW_0(\lambda)}{\left\{\int_0^1 [K_{c_1,c_2}(\lambda)]^2 d\lambda\right\}^{1/2}} \tag{21}$$

$$K_{c_1,c_2}(\lambda) = J_{c_2}^{\mu}(\lambda) - A J_{c_1}^{\mu}(\lambda)$$

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.[10] In other words, if x_{1t} is strictly exogenous ($\delta = 0$) then the OLS t-test of $\beta_2 = 0$ will be valid in small samples even with highly persistent regressors. By contrast, the term $dW_1(\lambda)$ in the numerator of (20) is not independent of the denominator and this gives Z_1 a nonstandard distribution. In particular, Appendix B establishes that $\text{Var}(Z_1) > 1$. Moreover Z_1 and Z_0 are uncorrelated with each other.[11] Therefore the t-statistic in (19) in general 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

[10] 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 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).

[11] 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.

when H_0 is true.

Expression (19) 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 $J^{\mu}_{c_1}(\lambda)$. 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 (19). Unlike the Dickey-Fuller distribution, Z_1 has mean zero, but like the Dickey-Fuller distribution it has variance larger than one.

2.2.2 Simulation evidence

We now examine the implications of the theory developed above in a simulation study. We generate values for x_{1t} and x_{2t} using (14), with ε_{1t} and ε_{2t} serially independent Gaussian random variables with unit variance and covariance equal to θ.[12] We then calculate

$$y_{t+1} = \rho_1 x_{1t} + u_{t+1}, \quad u_t = \delta \varepsilon_{1t} + \sqrt{1-\delta^2} v_t,$$

where v_t is an *i.i.d.* standard normal random variable. This implies that in the predictive equation (15) the true parameters are $\beta_0 = \beta_2 = 0$ and $\beta_1 = \rho_1$, and that the correlation between u_t and ε_{1t} is δ. Note that for $\delta = 1$ this corresponds to a model with a lagged dependent variable ($y_t = x_{1t}$), whereas for $\delta = 0$ both predictors are strictly exogenous as u_t is independent of both both ε_{1t} and ε_{2t}. While in bond return regressions δ is typically negative (as we discuss below in Section 3), we can focus here on $0 \leq \delta \leq 1$, since only $|\delta|$ matters for the distribution of the t-statistic.

We first set $\theta = 0$ as in our theory above, so that the variance matrix V is given by equation (16) with σ_1, σ_2, and σ_u equal to one, and x_{2t} is strictly exogenous. We investigate the effects of varying δ, the persistence of the predictors ($\rho_1 = \rho_2 = \rho$), and the sample size T. We simulate 50,000 artificial data samples, and in each sample we estimate the regression in equation (15). Since our interest is in the inference about β_2 we use this simulation design to study the small-sample behavior of the t-statistic for the test of $H_0 : \beta_2 = 0$. To give

[12] We start the simulations at $x_{1,0} = x_{2,0} = 0$, following standard practice of making all inference conditional on date 0 magnitudes.

conventional inference the best chance, we use OLS standard errors, which is the correct choice in this simulation setup as the errors are not serially correlated ($h = 1$) and there is no heteroskedasticity.[13]

In addition to the small-sample distribution of the t-statistic we also study its asymptotic distribution given in equation (19). While this is a non-standard distribution, we can draw from it using Monte Carlo simulation: for given values of c_1 and c_2, we simulate samples of size $\tilde{T}$ from near-integrated processes and approximating the integrals using Rieman sums—see, for example, Chan (1988), Stock (1991), and Stock (1994). The literature suggests that such a Monte Carlo approach yields accurate approximations to the limiting distribution even for moderate sample sizes (Stock, 1991, uses $\tilde{T} = 500$). We will use $\tilde{T} = 1000$ and generate 50,000 Monte Carlo replications with $c_1 = c_2 = T(\rho - 1)$ to calculate the predicted outcome for a sample of size T with serial dependence ρ.

Table 1 reports the performance of the t-test of H_0 with a nominal size of five percent. It shows the true size of this test, i.e., the frequency of rejections of H_0, according to both the small-sample distribution from our simulations and the asymptotic distribution in equation (19). We use critical values from a Student t-distribution with $T - 3$ degrees of freedom. Not surprisingly, the local-to-unity asymptotic distribution provides an excellent approximation to the exact small-sample distributions, as both indicate a very similar test size across parameter configurations and sample sizes. The main finding here is that the size distortions can be quite substantial with a true size of up to 17 percent—the t-test would reject the null more than three times as often as it should. When $\delta \neq 0$, the size of the t-test increases with the persistence of the regressors. Table 1 also shows the dependence of the size distortion on the sample size. To visualize this, Figure 1 plots the empirical size of the t-test for the case with $\delta = 1$ for different sample sizes from $T = 50$ to $T = 1000$.[14] When $\rho < 1$, the size distortions decrease 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 (19) with $c_i = 0$ governs the distribution for arbitrarily large T.

To understand better why conventional t-tests go so wrong in this setting, we use simulations to study the respective roles of bias in the coefficient estimates and of inaccuracy of the OLS standard errors for estimation of β_1 and β_2. Table 2 shows results for three different simulation settings, in all of which $T = 100$, $\rho = 0.99$, and x_{1t} is correlated with past forecast errors ($\delta \neq 0$). In the first two settings, the correlation between the regressors is zero ($\theta = 0$),

[13] If we instead use Newey-West standard errors, the size distortions become larger, as expected based on the well-known small-sample problems of HAC covariance estimators (e.g., Müller, 2014).

[14] The lines in Figure 1 are based on 500,000 simulated samples in each case.

and δ is either equal to one or 0.8. In the third setting, we investigate the effects of non-zero correlation between the predictors by setting $\delta = 0.8$ and $\theta = 0.8$.[15] The results show that in all three simulation settings b_1 is downward biased and b_2 is unbiased. The problem with the hypothesis test of $\beta_2 = 0$ does not arise from coefficient (Stambaugh) bias, but from the fact that the asymptotic standard errors underestimate the true sampling variability of both b_1 and b_2, i.e., from "standard error bias." This is evident from comparing the standard deviation of the coefficient estimates across simulations—the true small-sample standard error—and the average OLS standard errors. The latter are between 22 and 31 percent too low. Because of this standard error bias, the tests for $\beta_2 = 0$ reject much more often than their nominal size of five percent.

2.2.3 Relevance for tests of the spanning hypothesis

We have demonstrated that with persistent predictors, the lack of strict exogeneity of a subset of the predictors can have serious consequences for the small-sample inference on the remaining predictors, because it causes standard error bias for all predictors. Importantly, HAC standard errors do not help, because in such settings they cannot accurately capture the uncertainty surrounding the coefficient estimators. This econometric issue arises necessarily in all tests of the spanning hypothesis. First, in these regressions the predictors in x_{1t} are by construction correlated with u_t, because they correspond to information in current yields and the dependent variable is a future bond return. Second, the predictors are often highly persistent. Table 3, which we discuss in more detail below, reports the estimated autocorrelation coefficients for the predictors used in each published study, showing the high persistence of the predictors used in practice. Third, the sample sizes are necessarily small.[16] In light of these observations, conventional hypothesis tests are likely to be misleading in all of the empirical studies that we consider in this paper.

Predictive regressions for bond returns are "unbalanced" in the sense that the dependent variable has little serial correlation whereas the predictors are highly persistent. One might suppose that inclusion of additional lags solves the problem we point out. This, unfortunately, is not the case: including further lags of x_{2t} and testing whether the coefficients on current and lagged values are jointly significant leads to a test with exactly the same small-sample

[15] Note that in this setting, x_{2t} is not strictly exogenous, as the correlation between u_t and ε_{2t} is $\theta\delta$. This is the natural implication of a model in which only x_{1t} contains information useful for predicting y_t. If instead we insisted on $E(u_t\varepsilon_{2t}) = 0$ while $\theta \neq 0$ (or, more generally, if $E(u_t\varepsilon_{2t}) \neq \theta\delta$) then $E(y_t|x_{1t}, x_{1,t-1}, x_{2t}, x_{2,t-1}) \neq E(y_t|x_{1t}, x_{1,t-1})$ meaning that in effect y_t would depend on both x_{1t} and x_{2t}.

[16] Reliable interest rate data are only available since about the 1960s, which leads to situations with about 40-50 years of monthly data. Going to higher frequencies—such as weekly or daily—does not increase the effective sample sizes, since it typically increases the persistence of the series and introduces additional noise.

size distortions as the t-test on x_{2t} alone.[17]

2.3 A bootstrap design for investigating the spanning hypothesis

The above analysis suggests that it is of paramount importance to base inference on the small-sample distributions of the relevant test statistics. We propose to do so using a parametric bootstrap under the spanning hypothesis.[18] While some studies (Bekaert et al., 1997; Cochrane and Piazzesi, 2005; Ludvigson and Ng, 2009; Greenwood and Vayanos, 2014) use the bootstrap in a similar context, they typically generate data under the expectations hypothesis. Cochrane and Piazzesi (2005) and Ludvigson and Ng (2009, 2010) also calculated bootstrap confidence intervals under the alternative hypothesis, which in principle gives some indication of the small-sample significance of the coefficients on x_{2t}. However, bootstrapping under the relevant null hypothesis—the spanning hypothesis—is much to be preferred, as it allows us to calculate the small-sample size of conventional tests and generally leads to better numerical accuracy and more powerful tests (Hall and Wilson, 1991; Horowitz, 2001). Our paper is the first to propose a bootstrap to test the spanning hypothesis $H_0 : \beta_2 = 0$ by generating bootstrapped samples under the null.

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:

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

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 $\hat{W}\hat{W}' = I_3$.[19] Fitted yields can be obtained using $\hat{\imath}_t = \hat{W}' x_{1t}$. 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.[20]

[17] A closely related problem arises in classical spurious regression, see Hamilton (1994, p. 562))

[18] 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.

[19] We choose the eigenvectors so that the elements in the last column of $\hat{W}$ are positive—see also footnote 7.

[20] For example, in the case study of Joslin et al. (2014) in Section 3, the standard deviation is 6.5 basis points.

Then we estimate by OLS a VAR(1) for x_{1t}:

$$x_{1t} = \hat{\phi}_0 + \hat{\phi}_1 x_{1,t-1} + e_{1t} \qquad t = 1, \ldots, T. \tag{22}$$

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. Studies that have documented that such a simple factor model fits and forecasts the yield curve well include Duffee (2011) 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[21] on

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

where $e^*_{1\tau}$ denotes bootstrap residuals. Then we obtain the artificial yields using

$$i^*_{n\tau} = \hat{w}'_n x^*_{1\tau} + v^*_{n\tau} \tag{23}$$

for $v^*_{n\tau} \sim N(0, \sigma_v^2)$. The standard deviation of the measurement errors, σ_v, is set to the sample standard deviation of the fitting errors $\hat{v}_{nt}$.[22]

We thus have generated an artificial sample of yields $i^*_{n\tau}$ which by construction only three factors (the elements of $x^*_{1\tau}$) have any power to predict, but whose covariance and dynamics are similar to those of the observed data i_{nt}. Notably, our bootstrapped yields are first-order Markov—under our bootstrap the current yield curve contains all the information necessary to forecast future yields.

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

$$x_{2t} = \hat{\alpha}_0 + \hat{\alpha}_1 x_{2,t-1} + e_{2t}, \tag{24}$$

from which we then bootstrap 5000 artificial samples $x^*_{2\tau}$ in a similar fashion as for $x^*_{1\tau}$. The bootstrap residuals $(e'^*_{1\tau}, e'^*_{2\tau})$ are drawn from the joint empirical distribution of (e'_{1t}, e'_{2t}).[23]

[21] We start the recursion with a draw from the unconditional distribution implied by the estimated VAR for x_{1t}.

[22] We can safely assume serially uncorrelated fitting errors, despite some evidence in the literature to the contrary (Adrian et al., 2013; Hamilton and Wu, 2014). Recall that our goal is to investigate the small-sample properties of previously calculated test statistics in an environment in which the null hypothesis holds by construction. Adding serial correlation in $v^*_{n\tau}$ would only add yet another possible reason why the spanning hypothesis could have been spuriously rejected by earlier researchers.

[23] We also experimented with a Monte Carlo design in which $e^*_{1\tau}$ was drawn from a Student-t dynamic

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 for which the dynamic serial correlation of yields and explanatory variables are similar to those in the actual data but in which by construction the null hypothesis is true that $x^*_{2\tau}$ has no predictive power for future yields and bond returns.[24] In particular, under our bootstrap there are no unspanned macro risks. To see how to test the spanning hypothesis using the bootstrap, 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. We calculate the bootstrap p-value as the fraction of samples in which $|t^*_i| > |t|$, and would reject the null if this is less than, say, five percent. In addition, we can estimate the true size of a conventional t-test as the fraction of samples in which $|t^*_i|$ exceeds the usual asymptotic critical value.

One concern about this procedure is related to the well-known fact that under local-to-unity asymptotics, the bootstrap generally cannot provide a test of the correct nominal size.[25] The reason is that the test statistics are not asymptotically pivotal as their distribution depends on the nuisance parameters c_1 and c_2, which cannot be consistently estimated. For our purpose, however, this is not a concern for two reasons. First, when the goal (as in this investigation) is to judge whether the existing evidence against the spanning hypothesis is compelling, we do not need to be worried about a test that is not conservative enough. Let's say our bootstrap procedure does not completely eliminate the size distortions and leads to a test that still rejects somewhat too often. If such a test nevertheless fails to reject the spanning hypothesis, we know this could not be attributed to the test being too conservative, but instead accurately conveys a lack of evidence against the null. Nor is a failure to reject a reflection of a lack of power. In additional, unreported results we have found that for those coefficients that are non-zero in our bootstrap DGP, we consistently and strongly reject the null.

Moreover, we can directly evaluate the accuracy of our bootstrap procedure using simula-

conditional correlation GARCH model (Engle, 2002) fit to the residuals e_{1t} with similar results to those obtained using independently resampled e_{1t} and e_{2t}.

[24] For example, if y_{t+h} is an h-period excess return as in equation (8) then in our bootstrap

$$\begin{aligned} y^*_{\tau+h} &= ni^*_{n\tau} - hi^*_{h\tau} - (n-h)i^*_{n-h,\tau+h} \\ &= n(\hat{w}'_n x^*_{1\tau} + v^*_{n\tau}) - h(\hat{w}'_h x^*_{1\tau} + v^*_{h\tau}) - (n-h)(\hat{w}'_{n-h} x^*_{1,\tau+h} + v^*_{n-h,\tau+h}) \\ &= n(\hat{w}'_n x^*_{1\tau} + v^*_{n\tau}) - h(\hat{w}'_h x^*_{1\tau} + v^*_{h\tau}) - (n-h)[\hat{w}'_{n-h}(\hat{k}_h + e^*_{1,\tau+h} + \hat{\phi}_1 e^*_{1,\tau+h-1} + \cdots \\ &+ \hat{\phi}_1^{h-1} e^*_{1,\tau+1} + \hat{\phi}_1^h x^*_{1\tau}) + v^*_{n-h,\tau+h}] \end{aligned}$$

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

[25] This result goes back to Basawa et al. (1991). See also Hansen (1999) as well as Horowitz (2001) and the references therein.

tions. It is straightforward to use the Monte Carlo simulations in Section 2.2.2 to calculate what the size of our bootstrap procedure would be if applied to a specified parametric model. In each sample i simulated from a known parametric model, we can: (i) calculate the usual 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 simulation using these estimated autoregressive coefficients; (iv) estimate the predictive regression on the bootstrap simulation;[26] (v) calculate the t-test of $\beta_2 = 0$ on this bootstrap predictive regression, denoted t_i^*. We generate 5,000 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 second-to-last row of Table 2. We find in these settings that our bootstrap has a size above but fairly close to five percent. The size distortion is always smaller for our bootstrap than for the conventional t-test.

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. In addition, we find that our bootstrap procedure has good power properties. The implication is that if our bootstrap procedure fails to reject the spanning hypothesis, we can safely conclude that the evidence against the spanning hypothesis in the original data is not persuasive.

A separate but related issue is that least squares typically underestimates the autocorrelation of highly persistent processes due to small-sample bias (Kendall, 1954; Pope, 1990). Therefore the VAR we use in our bootstrap would typically be less persistent than the true data-generating process. For this reason, we might expect the bootstrap procedure to be slightly oversized.[27] One way to deal with this issue is to generate samples not from the OLS estimates $\hat{\phi}_1$ and $\hat{\alpha}_1$ but instead use bias-corrected VAR estimates obtained with the bootstrap adopted by Kilian (1998). We refer to this below as the "bias-corrected bootstrap."[28]

[26] 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 all our empirical analysis the bootstrapped dependent variable is obtained from (23) and the definition of y_{t+h} (for example, equation (8)).

[27] A test that would have size five percent if the serial correlation was given by $\hat{\rho}_1 = 0.97$ would have size greater than five percent if the true serial correlation is $\rho_1 = 0.99$.

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

2.4 An alternative robust test for predictability

There is of course a very large literature addressing the problem of HAC inference. This literature is concerned with accurately estimating the matrix S in (7) but does not address what we have identified as the key issue, which is the small-sample difference between the statistics in (9) and (12). We have looked at a number of alternative approaches in terms of how well they perform in our bootstrap experiments. We found that the most reliable existing test appears to be the one suggested by Ibragimov and Müller (2010), who proposed a novel method for testing a hypothesis about a scalar coefficient. The original dataset is divided into q subsamples and the statistic is estimated separately over each subsample. If these estimates across subsamples are approximately independent and Gaussian, then a standard t-test with q degrees of freedom can be carried out to test hypotheses about the parameter. Müller (2014) provided evidence that this test has excellent size and power properties in regression settings where standard HAC inference is seriously distorted. Our simulation results, to be discussed below, show that this test also performs very well in the specific settings that we consider in this paper, namely inference about predictive power of certain variables for future interest rates and excess bond returns. Throughout this paper, we report two sets of results for the Ibragimov-Müller (IM) test, setting the number of subsamples q equal to either 8 and 16 (as in Müller, 2014). A notable feature of the IM test is that it allows us to carry out inference that is robust not only with respect to serial correlation but also with respect to parameter instability across subsamples, as we will discuss Section 5.

We use the same Monte Carlo simulation as before to estimate the size of the IM test in the simple setting with two scalar predictors. The results are shown in the last row of Table 2. The IM test has close to nominal size in all three settings. The reason is that the IM test is based on more accurate estimates of the sampling variability of the test statistic by using variation across subsamples. In this way, it solves the problem of standard error bias that conventional t-tests are faced with. Note, however, that coefficient bias would be a problem for the IM test, because it splits the (already small) sample into even smaller samples, which would magnify the small-sample coefficient bias. It is therefore important to assess whether the conditions are met for the IM test to work well in practice, which we will do below in our empirical applications. It will turn out that in our applications the IM test should perform very well.

percent than for the simple bootstrap.

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, all of those results stem from the substantial in-sample predictive power of x_{2t} in these excess return regressions. The sample contains monthly observations over the period 1985:1-2007:12.

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

JPS found that for the ten-year bond, the adjusted $\bar{R}^2$ of regression (1) when x_{2t} is excluded is only 0.20. But when they added x_{2t}, the $\bar{R}^2$ increased to 0.37. For the two-year bond, the change is even more striking, with $\bar{R}^2$ increasing from 0.14 without the macro variables to 0.48 when they are included. JPS interpreted these adjusted $\bar{R}^2$ as strong evidence that macroeconomic variables have predictive power for excess bond returns beyond the information in the yield curve itself, and concluded from this evidence that "macroeconomic risks are unspanned by bond yields" (p. 1203).

However, there are some warning flags for these predictive regressions, which we report in Table 3. First, the predictors in x_{2t} are very persistent. The first-order sample autocorrelations for GRO and INF are 0.91 and 0.99, respectively. The yield PCs in x_{1t}, in particular the level and slope, are of course highly persistent as well, which is a common feature of interest rate data. Second, to assess strict exogeneity of the predictors we report estimated values for δ, the correlation between innovations to the predictors, ε_{1t} and ε_{2t}, and the lagged prediction error, u_t.[29] The innovations are obtained from the estimated VAR models for x_{1t} and x_{2t}, and the prediction error is calculated from least squares estimates of equation (1) for y_{t+h} the average excess bond return for two- through ten-year maturities. For the first PC of yields, the level of the yield curve, strict exogeneity is strongly violated, as the absolute value of δ is substantial. Its sizable negative value is due to the mechanical relationship between bond returns and the level of the yield curve: a positive innovation to PC1 at t raises all yields and mechanically

[29] While in our theory in Section 2.2 δ was the correlation of the (scalar) innovation of x_{1t} with past prediction errors, here we calculate it for all predictors in x_{1t} and x_{2t}.

lowers bond returns from $t-h$ to t. Hence such a violation of strict exogeneity will always be present in predictive regressions for bond returns that include the current level of the yield curve. In light of our results in Section 2 these warning flags suggest that small-sample issues are present, and we will use the bootstrap to address them.

Table 4 shows $\bar{R}^2$ of predictive regressions for the excess bond returns on the two- and ten-year bond, and for the average excess return across maturities. The first three columns are for the same data set as was used by JPS.[30] The first row in each panel reports the actual $\bar{R}^2$, and for the excess returns on the 2-year and 10-year bonds essentially replicates the results in JPS.[31] The entry $\bar{R}^2_1$ gives the adjusted $\bar{R}^2$ 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. The second row reports the mean $\bar{R}^2$ across 5000 replications of the bootstrap described in Section 2.3, 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% confidence intervals, calculated from the bootstrap distribution of the test statistics.

For all predictive regressions, the variability of the adjusted $\bar{R}^2$ is very high. Values for $\bar{R}^2_2$ up to about 63% would not be uncommon, as indicated by the bootstrap confidence intervals. Most notably, adding the regressors x_{2t} often substantially increases the adjusted R^2, by up to 23 percentage points or more, although x_{2t} has no predictive power in population by construction. For the ten-year bond, JPS report an increase of 17 percentage points when adding macro variables, but our results show that this increase is in fact not statistically significant at conventional significance levels. For the two-year bond, the increase in $\bar{R}^2$ of 35 percentage points is statistically significant. However, the two-year bond seems to be special among the yields one could look at. When we look for example at the average excess return across all maturities, our bootstrap finds no evidence that x_{2t} has predictive power beyond the information in the yield curve, as reported in the last panel of Table 4.

Since the persistence of x_{2t} is high, it may be important to adjust for small-sample bias in the VAR estimates. For this reason we also carried out the bias-corrected (BC) bootstrap. The expected values and 95% confidence intervals are reported in the bottom two rows of

[30] Their yield data set ends in 2008, with the last observation in their regression corresponding to the excess bond return from 2007:12 to 2008:12.

[31] The yield data set of JPS includes the six-month and the one- through ten-year Treasury yields. After calculating annual returns for the two- to ten-year bonds, JPS discarded the six, eight, and nine-year yields before fitting PCs and their term structure models. Here, we need the fitted nine-year yield to construct the return on the ten-year bond, so we keep all 11 yield maturities. While our PCs are therefore slightly different than those in JPS, the only noticeable difference is that our adjusted $\bar{R}^2$ in the regressions for the two-year bond with yield PCs and macro variables is 0.49 instead of their 0.48.

each panel of Table 4. As expected, more serial correlation in the generated data (due to the bias correction) increases the mean and the variability of the adjusted $\bar{R}^2$ and of their difference. In particular, while the difference $\bar{R}^2_2 - \bar{R}^2_1$ for the average excess return regression was marginally significant at the 10-percent level for the simple bootstrap, it is insignificant at this level for the BC bootstrap.

The right half of Table 4 updates the analysis to include an additional 7 years of data. As expected under the spanning hypothesis, the value of $\bar{R}^2_2$ that is observed in the data falls significantly when new data are added. And although the bootstrap 95% confidence intervals for $\bar{R}^2_1 - \bar{R}^2_2$ are somewhat tighter with the longer data set, the conclusion that there is no statistically significant evidence of added predictability provided by x_{2t} is even more compelling. For all bond maturities, the increases in adjusted $\bar{R}^2$ from adding macro variables as predictors for excess returns lie comfortably inside the bootstrap confidence intervals.

3.2 Testing the spanning hypothesis

Is the predictive power of macro variables statistically significant? JPS only reported adjusted $\bar{R}^2$ for their excess return regression, but one is naturally interested in formal tests of the spanning hypothesis in JPS' excess return regressions. The common approach to address the serial correlation in the residuals due to overlapping observations is to use the HAC standard errors and test statistics proposed by Newey and West (1987), typically using 18 lags (see among many others Cochrane and Piazzesi, 2005; Ludvigson and Ng, 2009). In the second row of Table 5 we report the resulting t-statistic for each coefficient along with the Wald test of the hypothesis $\beta_2 = 0$, calculated using Newey-West standard errors with 18 lags. The third row reports asymptotic p-values for these statistics. According to this popular test, GRO and INF appear strongly significant, both individually and jointly. In particular, the Wald statistic has a p-value below 0.1%.

We then employ our bootstrap to carry out tests of the spanning hypothesis that account for the small-sample issues described in Section 2. Again, we use both a simple bootstrap based on OLS estimates of the VAR parameters, as well as a bias-corrected (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 the simple bootstrap, the coefficient on GRO is insignificant, while INF is still marginally significant. Using the BC bootstrap, however, the coefficients are both individually and jointly insignificant, in stark contrast to the conventional HAC tests.

We also report in Table 5 the p-values for the IM test of the individual significance of the

coefficients. The level and slope of the yield curve ($PC1$ and $PC2$) are strongly significant predictors according to both IM tests.[32] This will turn out to be a consistent finding in all the data sets that we will look at—the level and slope of the yield curve appear to be robust predictors of bond returns, consistent with an old literature going back to Fama and Bliss (1987) and Campbell and Shiller (1991).[33] By contrast, the coefficients on GRO and INF are not statistically significant at conventional significance levels based on the IM test.

We then use the bootstrap to calculate the properties of the different tests for data with serial correlation properties similar to those observed in the sample. In particular, we estimate the true size of the HAC, bootstrap, and IM tests with nominal size of five percent, and report these in the last four rows of the top panel of Table 5. For the HAC tests, this is simply the frequency of bootstrap replications in which the t- and Wald-statistics exceed the usual asymptotic critical values. The results reveal that the true size of the conventional tests is 21-38% instead of the presumed five percent.[34] These substantial size distortions are also reflected in the bootstrap critical values, which far exceed the conventional critical values. The bootstrap and the IM tests, in contrast, have a size that is estimated to be very close to five percent, eliminating almost all of the size distortions of the more conventional tests.

As in the originally published work, we study returns with twelve-month holding periods in all empirical applications of this paper. One might be interested, however, in the magnitude of the size distortions for one-month bond returns. In such a setting, only the lack of strict exogeneity of x_{1t} causes problems for small-sample inference, and not the serial correlation in the prediction errors. In additional, unreported results using the JPS data, we find that in regressions for one-month excess returns the bootstrap does not reject the spanning hypothesis. The conventional tests have serious size distortions, which are however smaller than in the presence of serially correlated errors.[35] The implication is that the substantial small-sample size distortions we reported above for data with overlapping returns are due to a combination of both problems, serially correlated errors as well as lack of strict exogeneity.

When we add more data to the sample, we again find that the statistical evidence of predictability declines substantially, as seen in the second panel of Table 5. When the data set is extended through 2013, the HAC test statistics are only marginally significant or insignificant, even if interpreted assuming the usual asymptotics. Using the bootstrap to take into account

[32] The low p-values are also consistent with the conclusion from our unreported Monte Carlo investigation that IM has good power to reject a false null hypothesis.

[33] We have also calculated small-sample confidence intervals using the bootstrap, which confirm that the coefficients on $PC1$ and $PC2$ are significant.

[34] Using the BC bootstrap gives an even higher estimate of the true size of the HAC Wald test, about 45%.

[35] Specifically, if we use White standard errors, as Duffee (2013b) and others do for predictions of one-month excess returns, the BC bootstrap estimate of the true size of the Wald test of the spanning hypothesis is 15%.

the small-sample size distortions of such tests, these test statistics are far from significant. Regarding the results for the IM test, we also find in this extended sample that the slope is an important predictor of excess bond returns, consistent with a large existing literature, whereas the coefficients on the macro variables are insignificant.

We conclude that the evidence in JPS on the predictive power of macro variables for yields and bond returns is not altogether convincing. Notwithstanding, JPS noted that theirs is only one of several papers claiming to have found such evidence. We turn to these studies in the following sections.

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, adding further evidence for the claim of unspanned macro risks and against the hypothesis of invertibility. Here we revisit this evidence, focusing on the results in Ludvigson and Ng (2010) (henceforth LN).

LN 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. The authors carried out an extensive specification search in which they considered many different combinations of the factors along with squared and cubic terms. They also included in their specification search the bond-pricing factor proposed by Cochrane and Piazzesi (2005), which is the linear combination of forward rates that best predicts the average excess return across maturities, and which we denote here by CP. LN's conclusion was that macro factors appear to help predict excess returns, even when controlling for the CP factor. This conclusion is mostly based on comparisons of adjusted $\bar{R}^2$ in regressions with and without the macro factors and on HAC inference using Newey-West standard errors.

4.1 Robust inference about coefficients on macro factors

One feature of LN's design obscures the evidence relevant for the null hypothesis that is the focus of our paper. Their null hypothesis is that the CP factor alone provides all the information necessary to predict bond yields, whereas our null hypothesis of interest is that the 3 variables $(PC1, PC2, PC3)$ contain all the necessary information. Their regressions in which CP alone is used to summarize the information in the yield curve could not be used to

test our null hypothesis. For this reason, we begin by examining similar predictive regressions to those in LN in which excess bond returns are regressed on three PCs of the yields and all eight of the LN macro factors. We further leave aside the specification search of LN in order to focus squarely on hypothesis testing for a given regression specification.[36] These regressions take the same form as (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. As before, our interest is in testing the hypothesis $H_0 : \beta_2 = 0$.

The top panel of Table 6 reports regression results for LN's original sample. The first three rows show the coefficient estimates, HAC t- and Wald statistics (using Newey-West standard errors with 18 lags as in LN), and p-values based on the asymptotic distributions of these test statistics. 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 (the five-percent critical value for a χ^2_8-distribution is 15.5). Table 7 reports adjusted $\bar{R}^2$ for the restricted ($\bar{R}^2_1$) and unrestricted ($\bar{R}^2_2$) regressions, and shows that this measure of fit increases by 10 percentage points when the macro factors are included. Taken at face value, this evidence suggests that macro factors have strong predictive power, above and beyond the information contained in the yield curve, consistent with the overall conclusions of LN.

How robust are these econometric results? We first check the warning flags summarized in Table 3. As usual, the yield PCs are very persistent. The macro factors differ in their persistence, but even the most persistent ones only 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. Again the first PC of yields strongly violates strict exogeneity for the reasons explained above. Based on these indicators, it appears that small-sample problems may well distort the results of conventional inference methods.

To assess the potential importance in this context, we 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.

Before turning to the results, it is worth noting the differences between our bootstrap exercise and the bootstrap carried out by LN. Their bootstrap is designed to test the null hypothesis that excess returns are not predictable against the alternative that they are pre-

[36] 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.

dictable by macro factors and the CP factor. Using this setting, LN produced convincing evidence that excess returns are predictable, which is fully consistent with all the results in our paper as well. 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 allows us to provide a more accurate assessment of the spanning hypothesis, and to estimate the size of conventional tests under the null.

As seen in Table 6, 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. Table 7 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 confidence interval.

Table 6 also reports p-values for the two IM tests using $q = 8$ and 16 subsamples. Only the coefficient on $F7$ is significant at the 5% level using this test, and then only for $q = 16$. The robustly significant predictors are the level and the slope of the yield curve.

We again use the bootstrap to estimate the true size of the different tests with a nominal size of five percent. The results, which are reported in the bottom four rows of the top panel of Table 6, reveal that the conventional tests have serious size distortions. The true size of these t-tests is 9-14 percent, instead of the nominal five percent, and for the Wald test the size distortion is particularly high with a true size of 34 percent. By contrast, the bootstrap and IM tests, according to our calculations, appear to have close to correct size.

The failure to reject the null based on the IM tests is a reflection of the fact that the parameter estimates are often unstable across subsamples. 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. To explore this further we repeated our analysis using the same 1985-2013 sample period that was used in the second panel of Tables 4 and 5. Note that whereas in the case of JPS this was a strictly larger sample than the original, in the case of LN our second 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.[37] We used the macro data set of McCracken and Ng (2014), to extract macro factors in the same way as LN over

[37] We also analyzed the full 1964-2013 sample and obtained similar results as over the 1964-2007 sample.

the more recent data.[38]

The bottom panels of Tables 6 and 7 display the results. Over the later sample period, the evidence for the predictive power of macro factors is even weaker. Notably, the Wald tests reject H_0 for both bond maturities (at the ten-percent level for the five-year bond) when using asymptotic critical values, but are very far from significant when using bootstrap critical values. The increases in adjusted $\bar{R}^2$ in Table 7 are not statistically significant, and the IM tests find essentially no evidence of predictive power of the macro factors.

These results imply that the evidence that macro factors have predictive power beyond the information already contained in yields is much weaker than the results in LN would initially have suggested. For the original sample used by LN, our bootstrap procedure reveals substantial small-sample size distortions and weakens the statistical significance of the predictive power of macro variables, while the IM test indicates that only the level and slope are robust predictors. For the later sample, there is no evidence for unspanned macro risks at all. Our overall conclusion is that the predictive power of macro variables is much more tenuous than one would have thought from the published results, and that both small-sample concerns as well as subsample stability raise serious robustness concerns.

4.2 Robust inference about return-forecasting factors

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 $H8$. The CP factor of Cochrane and Piazzesi (2005) is constructed similarly using a regression on five forward rates. Adding $H8$ to a predictive regression with CP substantially increases the adjusted $\bar{R}^2$, 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 8 and 9 replicate LN's results for these regressions on the macro- ($H8$) and yield-based (CP) return-forecasting factors.[39] Table 8 shows coefficient estimates and statistical significance, while Table 9 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 adjusted $\bar{R}^2$ by 9-11 percentage points.

[38] Using this macro data set and the same sample period as LN we obtained results that were very similar to those in the original paper, which gives us confidence in the consistency of the macro data set.

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

How plausible would it have been to obtain these results if macro factors have in fact no predictive power? In order to answer this question, we adjust our bootstrap design to handle regressions with return-forecasting factors CP and $H8$. 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. The results in Table 8 show that the bootstrap p-values are substantially larger than the asymptotic HAC p-values, and $H8$ is no longer significant at the 1% level. Table 9 shows that the observed increases in adjusted $\bar{R}^2$ when adding $H8$ to the regression are not statistically significant at the five-percent level, with the exception of the two-year bond maturity where the observed value lies slightly outside the 95% bootstrap confidence interval.

We report bootstrap estimates of the true size of conventional HAC tests and of our bootstrap test of the significance of the macro return-forecasting factor—for a nominal size of five percent—in the bottom two rows of the top panel of Table 8. 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 50-55 percent. In contrast, the size of our bootstrap test is estimated to be very close to the nominal size.

We also examined the same regressions over the 1985–2013 sample period with results shown in the bottom panel of Table 8 and in the right half of Table 9. In this sample, the return-forecasting factors would again both 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. The observed increases in $\bar{R}^2$ are squarely in line with what we would expect under the spanning hypothesis, as indicated by the confidence intervals in Table 9.

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—Table 3 shows that both $H8$ and CP have autocorrelations that are near 0.8 at first order, 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 uncertainty not accounted for by conventional inference. In such a setting other econometric methods—preferably a bootstrap exercise designed to assess the relevant null hypothesis—are needed to accurately carry out inference. For the case at hand, we conclude that a return-forecasting factor based on macro factors exhibits only very tenuous predictive power, much weaker than indicated by LN's original analysis and which disappears completely over a different sample period.

5 Higher-order PCs of yields

Cochrane and Piazzesi (2005) (henceforth CP) documented several striking new facts about excess bond returns. Focusing on returns with a one-year holding period, they showed that the same linear combination of forward rates predicts excess returns on different long-term bonds, that the coefficients of this linear combination have a tent shape, and that predictive regressions using this one variable deliver 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 significant predictors of future bond returns (see CP's Table 4 on p. 147, row 3).

In CP's data, the first three PCs explain 99.97% of the variation in the five Fama-Bliss yields (see page 147 of CP), consistent with the long-standing evidence that three factors are sufficient to almost fully capture the shape and evolution of the yield curve, a result going back at least to Litterman and Scheinkman (1991). CP found that the other two PCs, which explain only 0.03% of the 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 predict yields and excess returns and that higher-order PCs do not contain additional predictive power.

The first 3 rows of Table 10 replicate the relevant results of CP using their original data. We estimate the predictive regression for the average excess bond return using five PCs as predictors, and carry out HAC inference in this model using Newey-West standard errors as in CP. The Wald statistic and R_1^2 and R_2^2 are identical to those reported by CP. 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. In contrast to the results found for JPS in Section 3 and LN in Section 4, our bootstrap finds that the CP results cannot be accounted for by small-sample size distortions. The main reason for this is that the 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, and falls far outside the 95% bootstrap interval under the null.

Interestingly, however, the IM tests would fail to reject the null hypothesis that $\beta_2 = 0$. These indicate that the coefficients on $PC4$ and $PC5$ are not statistically significant, and find only the level and slope to be robust predictors of excess bond returns. The bootstrap

estimates of the size of the IM test, reported in the bottom two rows of the top panel of Table 10, indicate that these tests have close to nominal size, giving us added reason to pay attention to these results.

Figure 2 provides some intuition about why the IM tests fail to reject. It shows the coefficients on each predictor across the $q = 8$ subsamples used in the IM test. The coefficients are standardized by dividing them by the sample standard deviation across the eight estimated coefficients for each predictor. Thus, the IM t-statistics, which are reported in the legend of Figure 2, are equal to the means of the standardized coefficients across subsamples, multiplied by $\sqrt{8}$. The figure shows that $PC1$ and $PC2$ had much more consistent predictive power across subsamples than $PC4$, whose coefficient switches signs several times. The strong association between $PC4$ and excess returns is mostly driven by the fifth subsample, which starts in September 1983 and ends in July 1988.[40] This illustrates that the IM test, which is designed to produce inference that is robust to serial correlation, at the same time delivers results that are robust to sub-sample instability. Only the level and slope have predictive power for excess bond returns in the CP data that is truly robust in both meanings of the word.

It is worth emphasizing the similarities and differences between the tests of interest to CP and in our own paper. Their central claim, with which we concur, is that the factor they have identified is a useful and stable predictor of bond returns. However, this factor is a function of all 5 PC's, and the first 3 of these account for 76% of the variation of the CP factor. Our claim is that it is the role of $PC1$-$PC3$ in the CP factor, and not the addition of $PC4$ and $PC5$, that makes the CP pricing factor turn out to be a useful and stable predictor of yields.

Thus our test for structural stability differs from those performed in CP and their accompanying online appendix. 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. Our tests, by contrast, look at stability of the role of each individual PC. We agree with CP that the first three PC's indeed have a stable predictive relation, as we confirmed with the IM tests in Table 10 and Figure 2, and in additional, unreported subsample analysis similar to that in CP's appendix. On the other hand, the predictive power of the 4th and 5th PC is much more tenuous, and is insignificant in most of the subsample periods that CP considered. Duffee (2013b, Section 7) also documented that extending CP's sample period to 1952–2010 alters some of their key results, and we have found that over Duffee's sample period the predictive power of higher-order PCs disappears.

In the bottom panel of Table 10 we report results for our preferred sample period, from

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

1985 to 2013. In this case, the coefficients on $PC4$ and $PC5$ are not significant for any method of inference, and the increase in R^2 due to inclusion of higher-order PCs are comfortably inside the 95% bootstrap intervals. At the same time, the predictive power of the level and slope of the yield curve is quite strong also in this sample. Although the standard HAC t-test fails to reject that the coefficient on the level is zero, the same test finds the coefficient on the slope to be significant, and the IM tests imply that both coefficients are significant.

Since CP used a sample period that ended more than ten years prior to the time of this writing, we can carry out a true out-of-sample test of our hypothesis of interest. We estimate the same predictive regressions as in CP, for excess returns on two- to five-year bonds as well as for the average excess return across bond maturities. The first two columns of Table 11 report the in-sample R^2 for the restricted models (using only $PC1$ to $PC3$) and unrestricted models (using all PCs). Then we construct expected future excess returns from these models using yield PCs[41] from 2003:1 through 2012:12, and compare these to realized excess returns for holding periods ending in 2004:1 through 2013:12. Table 11 shows the resulting root-mean-squared forecast errors (RMSEs). For all bond maturities, the model that leaves out $PC4$ and $PC5$ performs substantially better, with reductions of RMSEs around 20 percent. The test for equal forecast accuracy of Diebold and Mariano (1995) rejects the null, indicating that the performance gains of the restricted model are statistically significant. Figure 3 shows the forecast performance graphically, plotting the realized and predicted excess bond returns. Clearly, both models did not predict future bond returns very well, expecting mostly negative excess returns over a period when these turned out to be positive. In fact, the unconditional mean, estimated over the CP sample period, was a better predictor of future returns. This is evident both from Figure 3, which shows this mean as a horizontal line, and from the RMSEs in the last column of Table 11. Nevertheless, the unrestricted model implied expected excess returns that were more volatile and significantly farther off than those of the restricted model. Restricting the predictive model to use only the level, slope and curvature leads to more stable and more accurate return predictions.

We conclude from both our in-sample and out-of-sample results that the evidence for predictive power of higher-order factors is tenuous and sample-dependent. To estimate bond risk premia in a robust way, we recommend using only those predictors that consistently show a strong associations with excess bond returns, namely the level and the slope of the yield curve.

[41] PCs are calculated throughout using the loadings estimated over the original CP sample period.

6 Bond supply

In addition to macro-finance linkages, a separate literature studies the effects of the supply of bonds on prices and yields. The theoretical literature on the so-called portfolio balance approach to interest rate determination includes classic contributions going back to Tobin (1969) and Modigliani and Sutch (1966), as well as more recent work by Vayanos and Vila (2009) and King (2013). A number of empirical studies document the relation between bond supply and interest rates during both normal times and over the recent period of near-zero interest and central bank asset purchases, including Hamilton and Wu (2012), D'Amico and King (2013), and Greenwood and Vayanos (2014). Both theoretical and empirical work has convincingly demonstrated that bond supply is related to bond yields and returns.

However, our question here is whether measures of Treasury bond supply contain information that is not already reflected in the yield curve and that is useful for predicting future bond yields and returns. Is there evidence against the spanning hypothesis that involves measures of time variation in bond supply? At first glance, the answer seems to be yes. Greenwood and Vayanos (2014) (henceforth GV) found that 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 holds up to closer scrutiny. The sample period used in Greenwood and Vayanos (2014) is 1952 to 2008.[42]

To estimate the effects of bond supply on interest rates, GV estimate a broad variety of different regression specifications with yields and returns of various maturities as dependent variables. Here we are most interested in those regressions that control for the information in the yield curve. In the top panel of Table 12 we reproduce 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.[43] Like GV we use Newey-West standard errors with 36 lags.[44]

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 but is no longer significant at the conventional significance level of five percent when the yield spread is included in the regression, as seen in the second panel. But once again there are some

[42] As in JPS, the authors report a sample end date of 2007 but use yields up to 2008 to calculate one-year bond returns up to the end of 2007.

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

[44] There are small differences in our and their t-statistics that we cannot reconcile but which are unimportant for the results.

warning flags that raise doubts about the validity of HAC inference. Table 3 shows that the bond supply variable is extremely persistent—the first-order autocorrelation is 0.998—and the one-year yield and yield spread are of course highly persistent as well. This leads us to suspect that the true p-value likely exceeds the purported 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 do not apply our bootstrap procedure here because this bond return is not constructed from the observed yield curve.[45] Instead we rely on IM tests to carry out robust inference. Neither of the IM tests finds the coefficient on bond supply to be statistically significant. In contrast, the coefficient on the term spread is strongly significant for the HAC test and both IM tests.

We consider two additional regression specifications that are relevant in this context. The first controls for information in the yield curve by including, instead of a single term spread, the first three PCs of observed yields.[46] 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 12. Again, the coefficient on bond supply is only marginally significant for the HAC t-test, and insignificant for the IM tests. In contrast, the coefficients on both PC1 and PC2 are strongly significant for the IM tests.

Finally, we consider the most common specification where y_{t+h} is the one-year excess return, averaged across two- though five-year maturities. The last panel of Table 12 shows that in this case, the coefficient on bond supply is insignificant. Since Table 3 indicates that for this predictive regression both persistence as well as lack of strict exogeneity are warning flags, so we also apply our bootstrap procedure. We find that there is a significant size distortion for this hypothesis test, and the bootstrap p-value is substantially higher than the conventional p-value. There is robust evidence that PC1 and PC2 have predictive power for bond returns, as judged by the IM test, whereas this test indicates that bond supply is not a robust predictor.

Overall, 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.

[45] GV obtained this series from Ibbotson Associates.

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

7 Output gap

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).

We investigate the predictive regressions for excess bond returns y_{t+h} using the output gap at date $t-1$ (gap_{t-1}), measured as the deviation of the Fed's Industrial Production series from a quadratic time trend.[47] CPR lagged their measure by one month to account for the publication lag of the Fed's Industrial Production data. Table 13 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.[48] 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.[49] 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. However, the HAC t-statistics are only about a third of the OLS t-statistics, and indicate that the coefficient on gap is far from significant, with p-values above 20%.[50]

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.[51] One way to test the spanning hypothesis is to include CP instead of $\tilde{CP}$, for which we report the results in the third panel of Table 13. In this case, the coefficient on gap switches to a positive sign, and its Newey-West t-statistic remains insignificant. In contrast, both $\tilde{CP}$ and CP are strongly significant in these regressions.

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

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

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

[50] This indicates that CPR may have mistakenly reported the OLS instead of the Newey-West t-statistics

[51] 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).

Our preferred specification includes the first three PCs of the yield curve—see the last panel of Table 13. Importantly, the predictor *gap* is highly persistent, with a first-order autocorrelation coefficient of 0.975, as shown in Table 3, and the level PC is not strictly exogenous, so we need to worry about conventional t-tests to be substantially oversized. Hence we also include results for robust inference using the bootstrap and IM tests. 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. The IM tests do not reject the null.

Overall, there is no evidence that the output gap predicts bond returns. The level and in particular the slope of the yield curve, in contrast, are very strongly associated with future excess bond returns, in line with our finding throughout this paper.

8 Conclusion

The methods developed in our paper confirm a well established finding in the earlier literature– the current level and slope of the yield curve are robust predictors of future bond returns. That means that in order to test whether any other variables may also help predict bond returns, the regression needs to include the current level and slope, which are highly persistent lagged dependent variables. If other proposed predictors are also highly persistent, conventional tests of their statistical significance can have significant size distortions and the R^2 of the regression can increase dramatically when the variables are added to the regression even if they have no true explanatory power.

We proposed two strategies for dealing with this problem, the first of which is a simple bootstrap based on PCs and the second a robust t-test based on subsample estimates proposed by Ibragimov and Müller (2010). We used these methods to revisit five different widely cited studies, and found in each case that 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.

We emphasize that these results do not mean that fundamentals such as inflation, output, and bond supplies do not matter for interest rates. Instead, our conclusion is that any effects of these variables can be summarized in terms of the level, slope, and curvature. Once these three factors are included in predictive regressions, no other variables appear to have robust forecasting power for future yields or returns. Our results cast doubt on the claims for the existence of unspanned macro risks and support the view that it is not necessary to look

beyond the information in the yield curve to estimate risk premia in bond markets.

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.

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 (2015) "Resolving the Spanning Puzzle in Macro-Finance Term Structure Models," Working Paper 2015-01, Federal Reserve Bank of San Francisco.

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.

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

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.

Carrodus, Mark L and David EA Giles (1992) "The exact distribution of R 2 when the regression disturbances are autocorrelated," *Economics Letters*, Vol. 38, pp. 375–380.

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.

Chernov, Mikhail and Philippe Mueller (2012) "The Term Structure of Inflation Expectations," *Journal of Financial Economics*, Vol. 106, pp. 367–394.

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.

Coroneo, Laura, Domenico Giannone, and Michle Modugno (2015) "Unspanned Macroeconomic Factors in the Yields Curve," *Journal of Business and Economic Statistics*, p. forthcoming.

D'Amico, Stefania and Thomas B. King (2013) "Flow and stock effects of large-scale treasury purchases: Evidence on the importance of local supply," *Journal of Financial Economics*, Vol. 108, pp. 425–448.

Deng, Ai (2013) "Understanding Spurious Regression in Financial Economics," *Journal of Financial Econometrics*, pp. 1–29.

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

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

——— (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.

Engle, Robert (2002) "Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models," *Journal of Business & Economic Statistics*, Vol. 20, pp. 339–350.

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.

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

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.

Hansen, Bruce E (1999) "The grid bootstrap and the autoregressive model," *Review of Economics and Statistics*, Vol. 81, pp. 594–607.

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.

King, Thomas B. (2013) "A Portfolio-Balance Approach to the Nominal Term Structure," Working Paper 2013-18, Federal Reserve Bank of Chicago.

Koerts, Johannes and Adriaan Pieter Johannes Abrahamse (1969) *On the theory and application of the general linear model*: Rotterdam University Press Rotterdam.

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.

Modigliani, Franco and Richard Sutch (1966) "Innovations in interest rate policy," *The American Economic Review*, pp. 178–197.

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.

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 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.

Priebsch, Marcel (2014) "(Un)Conventional Monetary Policy and the Yield Curve," working paper, Federal Reserve Board, Washington, D.C.

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.

Swanson, Eric T (2015) "A macroeconomic model of equities and real, nominal, and defaultable debt," unpublished manuscript, University of California, Irvine.

Tobin, James (1969) "A general equilibrium approach to monetary theory," *Journal of money, credit and banking*, Vol. 1, pp. 15–29.

Vayanos, Dimitri and Jean-Luc Vila (2009) "A Preferred-Habitat Model of the Term Structure of Interest Rates," NBER Working Paper 15487, National Bureau of Economic Research.

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

Wright, Jonathan H. (2011) "Term Premia and Inflation Uncertainty: Empirical Evidence from an International Panel Dataset," *American Economic Review*, Vol. 101, pp. 1514–1534.

Appendix

A First-order asymptotic results

Here we provide details of the claims made in Section 2.1. 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 SSR 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

$$SSR_1 - SSR_2 = \sum (x_t' b - x_{1t}' b_1^*)^2. \tag{25}$$

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

$$x_t' b = x_{1t}' b_1^* + \tilde{x}_{2t}' b_2 \tag{26}$$

where $\tilde{x}_{2t}$ denotes the residuals from regressions of the elements of x_{2t} on x_{1t} and b_2 can be obtained from an OLS regression of $y_{t+h} - x_{1t}' b_1^*$ on $\tilde{x}_{2t}$.[52] Thus from (25) and (26),

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

If the true value of β_2 is zero, then by plugging (1) into the definition of b_2 and using the fact that $\sum \tilde{x}_{2t} x_{1t}' \beta_1 = 0$ (which follows from the orthogonality of $\tilde{x}_{2t}$ with x_{1t}) we see that

$$b_2 = \left(\sum \tilde{x}_{2t} \tilde{x}_{2t}'\right)^{-1} \left(\sum \tilde{x}_{2t} u_{t+h}\right) \tag{27}$$

$$\begin{aligned} SSR_1 - SSR_2 &= b_2' \left(\sum \tilde{x}_{2t} \tilde{x}_{2t}'\right) b_2 \\ &= \left(T^{-1/2} \sum u_{t+h} \tilde{x}_{2t}'\right) \left(T^{-1} \sum \tilde{x}_{2t} \tilde{x}_{2t}'\right)^{-1} \left(T^{-1/2} \sum \tilde{x}_{2t} u_{t+h}\right). \end{aligned} \tag{28}$$

[52] That is, $b_2 = \left(\sum \tilde{x}_{2t} \tilde{x}_{2t}'\right)^{-1} \left(\sum \tilde{x}_{2t} (y_{t+h} - x_{1t} b_1^*)\right)$ for $\tilde{x}_{2t}$ defined in (10) and (11). The easiest way to confirm the claim is to show that the residuals implied by (26) 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 orthogonality of $y_{t+h} - x_{1t}' b_1^* - \tilde{x}_{2t}' b_2$ to $\tilde{x}_{2t}$ follows directly from the definition of b_2. 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 $x_{2t} = \tilde{x}_{2t} + A_T x_{1t}$.

If x_t is stationary and ergodic, then it follows from the Law of Large Numbers that

$$T^{-1}\sum \tilde{x}_{2t}\tilde{x}'_{2t} = T^{-1}\sum x_{2t}x'_{2t} - \left(T^{-1}\sum x_{2t}x'_{1t}\right)\left(T^{-1}\sum x_{1t}x'_{1t}\right)^{-1}\left(T^{-1}\sum x_{1t}x'_{2t}\right)$$
$$\xrightarrow{p} E(x_{2t}x'_{2t}) - \left[E(x_{2t}x'_{1t})\right]\left[E(x_{1t}x'_{1t})\right]^{-1}\left[E(x_{1t}x'_{2t})\right]$$

which equals Q in (6) in the special case when $E(x_{2t}x'_{1t}) = 0$. For the last term in (28) we see from (10) and (11) that

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

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, so from the Central Limit Theorem

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

implying from (28) that

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

Thus from (3),

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

as claimed in (4).

Expression (27) also implies that

$$\sqrt{T}b_2 = \left(T^{-1}\sum \tilde{x}_{2t}\tilde{x}'_{2t}\right)^{-1}\left(T^{-1/2}\sum \tilde{x}_{2t}u_{t+h}\right) \xrightarrow{d} Q^{-1}r$$

from which (13) follows immediately.

B Local-to-unity asymptotic results

Here we provide details behind the claims made in Section 2.2. We know 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 (18) 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 (18). We also have from equation (2.17) in Stock (1994) that

$$T^{-1/2}x_{2,[T\lambda]} \Rightarrow \sigma_2 J_{c_2}(\lambda)$$

where $[T\lambda]$ denotes the largest integer less than $T\lambda$. From the Continuous Mapping Theorem,

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

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

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

$$T^{-2}\sum \tilde{x}^2_{2t} = \int_0^1 \{T^{-1/2}\tilde{x}_{2,[T\lambda]}\}^2 d\lambda \Rightarrow \sigma_2^2 \int_0^1 \{K_{c_1,c_2}(\lambda)\}^2 d\lambda. \tag{29}$$

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

$$\begin{aligned} T^{-1}\sum \tilde{x}_{2t}u_{t+1} &= T^{-1}\sum [x_{2t} - \bar{x}_2 - A_T(x_{1t} - \bar{x}_1)](\delta\sigma_u v_{1,t+1} + \sqrt{1-\delta^2}\sigma_u v_{0,t+1}) \\ &\Rightarrow \delta\sigma_2\sigma_u \int K_{c_1,c_2}dW_1 + \sqrt{1-\delta^2}\sigma_2\sigma_u \int K_{c_1,c_2}dW_0. \end{aligned} \tag{30}$$

Recalling (27), under the null hypothesis the t-test of $\beta_2 = 0$ can be written as

$$\tau = \frac{\sum \tilde{x}_{2t}u_{t+1}}{\{s^2\sum \tilde{x}^2_{2t}\}^{1/2}} = \frac{T^{-1}\sum \tilde{x}_{2t}u_{t+1}}{\{s^2T^{-2}\sum \tilde{x}^2_{2t}\}^{1/2}} \tag{31}$$

where

$$s^2 \xrightarrow{p} \sigma_u^2. \tag{32}$$

Substituting (32), (30), and (29) into (31) 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 (19).

Last we demonstrate that the variance of the variable Z_1 defined in (20) exceeds unity. We

can write

$$Z_1 = \frac{\int_0^1 J_{c_2}^{\mu}(\lambda) dW_1(\lambda)}{\left\{\int_0^1 [K_{c_1,c_2}(\lambda)]^2 d\lambda\right\}^{1/2}} - \frac{A\int_0^1 J_{c_1}^{\mu}(\lambda) dW_1(\lambda)}{\left\{\int_0^1 [K_{c_1,c_2}(\lambda)]^2 d\lambda\right\}^{1/2}} \tag{33}$$

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 (18). This means that the following inequality holds for all realizations:

$$\left| \frac{\int_0^1 J_{c_2}^{\mu}(\lambda) dW_1(\lambda)}{\left\{\int_0^1 [K_{c_1,c_2}(\lambda)]^2 d\lambda\right\}^{1/2}} \right| > \left| \frac{\int_0^1 J_{c_2}^{\mu}(\lambda) dW_1(\lambda)}{\left\{\int_0^1 [J_{c_2}^{\mu}(\lambda)]^2 d\lambda\right\}^{1/2}} \right|. \tag{34}$$

Adapting the argument made in footnote 10, the magnitude inside the absolute-value operator on the right side of (34) can be seen to have a $N(0,1)$ distribution. Inequality (34) thus establishes that the first term in (33) has a variance that is greater than unity. The second term in (33) 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.[53] In sum, these arguments show that $\text{Var}(Z_1) > 1$.

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

Table 1: Simulation study: size distortions of conventional t-test

		$\delta = 0$			$\delta = 0.8$			$\delta = 1$		
T		$\rho = 0.9$	0.99	1	0.9	0.99	1	0.9	0.99	1
50	simulated	5.1	4.9	5.1	8.1	11.1	11.4	10.2	15.1	15.9
50	asymptotic	4.5	4.4	4.6	8.4	11.0	11.5	10.5	14.9	15.4
100	simulated	4.8	5.1	5.2	7.1	11.4	12.2	8.4	15.2	16.2
100	asymptotic	4.5	4.7	4.8	7.0	11.1	11.9	8.4	15.0	16.0
200	simulated	5.0	5.1	5.0	6.1	11.1	12.4	6.8	14.5	16.5
200	asymptotic	4.9	4.9	4.9	6.2	10.7	12.0	7.2	14.6	16.6
500	simulated	5.0	5.0	5.0	5.4	8.9	12.2	5.7	11.6	17.0
500	asymptotic	5.0	4.8	4.9	5.4	9.2	12.3	5.8	11.6	16.9

True size (in percentage points) of a conventional t-test of $H_0 : \beta_2 = 0$ with nominal size of 5%, in simulated small samples and according to local-to-unity asymptotic distribution. δ determines the degree of endogeneity, i.e., the correlation of x_{1t} with the lagged error term u_t. The persistence of the predictors is $\rho_1 = \rho_2 = \rho$. For details on the simulation study refer to main text.

Table 2: Simulation study: coefficient bias and standard error bias

	$\delta = 1, \theta = 0$		$\delta = 0.8, \theta = 0$		$\delta = 0.8, \theta = 0.8$	
	β_1	β_2	β_1	β_2	β_1	β_2
True coefficient	0.990	0.000	0.990	0.000	0.990	0.000
Mean estimate	0.921	0.000	0.936	0.000	0.935	0.000
Coefficient bias	-0.069	0.000	-0.054	0.000	-0.055	0.000
True standard error	0.053	0.055	0.049	0.049	0.082	0.083
Mean OLS std. error	0.038	0.038	0.038	0.038	0.064	0.064
Standard error bias	-0.015	-0.017	-0.011	-0.011	-0.018	-0.019
Size of t-test		0.155		0.111		0.112
Size of bootstrap test		0.080		0.072		0.067
Size of IM test		0.047		0.047		0.045

Analysis of bias in estimated coefficients and standard errors for regressions in small samples with $T = 100$ and $\rho_1 = \rho_2 = 0.99$, as well as estimated size of conventional t-test, bootstrap, and IM tests. For details on the simulation study refer to main text.

Table 3: Warning flags for predictive regressions in published studies

Study	Predictor	ACF(l)			δ
		1	6	12	
JPS	PC1	0.974	0.840	0.696	-0.368
	PC2	0.973	0.774	0.467	-0.048
	PC3	0.849	0.380	0.216	0.202
	GRO	0.910	0.507	0.260	-0.122
	INF	0.986	0.897	0.815	-0.189
LN	PC1	0.984	0.904	0.821	-0.342
	PC2	0.944	0.734	0.537	0.137
	PC3	0.601	0.254	0.113	0.091
	F1	0.766	0.381	0.088	0.100
	F2	0.748	0.454	0.188	0.160
	F3	-0.233	0.035	-0.085	0.044
	F4	0.455	0.207	0.151	0.189
	F5	0.361	0.207	0.171	0.169
	F6	0.422	0.476	0.272	0.058
	F7	-0.111	0.134	0.054	-0.079
	F8	0.225	0.087	0.093	0.048
	CP	0.773	0.531	0.377	
	H8	0.777	0.627	0.331	
CP	PC1	0.980	0.880	0.767	-0.358
	PC2	0.940	0.721	0.539	0.157
	PC3	0.592	0.237	0.110	0.090
	PC4	0.425	0.137	0.062	-0.020
	PC5	0.227	0.157	-0.135	0.121
	CP	0.767	0.522	0.361	
GV	PC1	0.988	0.925	0.860	-0.312
	PC2	0.942	0.722	0.521	0.147
	PC3	0.582	0.233	0.094	0.105
	supply	0.998	0.990	0.974	0.035
CPR	PC1	0.986	0.917	0.841	-0.338
	PC2	0.939	0.712	0.528	0.179
	PC3	0.590	0.262	0.153	0.055
	gap	0.975	0.750	0.475	-0.193

Measures of persistence and lack of strict exogeneity of the predictors. For the persistence we report autocorrelations of the predictors at lags of one, six, and twelve months. Lack of strict exogeneity is measured by δ, the correlation between the innovations to the predictors, ε_{1t} or ε_{2t}, and the lagged prediction error, u_t. The innovations are obtained from estimated VAR(1) models for x_{1t} (the principal components of yields) and x_{2t} (the other predictors). The forecast error u_t is calculated from a predictive regression of the average excess bond return across maturities. The predictors are described in the main text. The data and sample are the same as in the published studies. These are JPS (Joslin et al., 2014), LN (Ludvigson and Ng, 2010), CP (Cochrane and Piazzesi, 2005), GV (Greenwood and Vayanos, 2014), and CPR (Cooper and Priestley, 2008).

Table 4: Joslin-Priebsch-Singleton: R^2 in excess return regressions

	Original sample: 1985–2008			Later sample: 1985–2013		
	$\bar{R}^2_1$	$\bar{R}^2_2$	$\bar{R}^2_2 - \bar{R}^2_1$	$\bar{R}^2_1$	$\bar{R}^2_2$	$\bar{R}^2_2 - \bar{R}^2_1$
Two-year bond						
Data	0.14	0.49	0.35	0.12	0.28	0.16
Simple bootstrap	0.30	0.36	0.06	0.26	0.32	0.06
	(0.06, 0.58)	(0.11, 0.63)	(-0.00, 0.22)	(0.05, 0.51)	(0.09, 0.56)	(-0.00, 0.21)
BC bootstrap	0.38	0.44	0.06	0.32	0.38	0.06
	(0.07, 0.72)	(0.13, 0.75)	(-0.00, 0.23)	(0.07, 0.60)	(0.12, 0.64)	(-0.00, 0.21)
Ten-year bond						
Data	0.20	0.37	0.17	0.20	0.28	0.08
Simple bootstrap	0.26	0.32	0.07	0.24	0.30	0.06
	(0.07, 0.48)	(0.12, 0.54)	(-0.00, 0.23)	(0.06, 0.46)	(0.11, 0.51)	(-0.00, 0.21)
BC bootstrap	0.27	0.34	0.08	0.26	0.33	0.07
	(0.06, 0.50)	(0.12, 0.57)	(-0.00, 0.27)	(0.06, 0.49)	(0.11, 0.55)	(-0.00, 0.23)
Average two- through ten-year bonds						
Data	0.19	0.39	0.20	0.17	0.25	0.08
Simple bootstrap	0.28	0.35	0.07	0.24	0.30	0.06
	(0.08, 0.50)	(0.12, 0.56)	(-0.00, 0.23)	(0.05, 0.46)	(0.10, 0.52)	(-0.00, 0.21)
BC bootstrap	0.30	0.37	0.07	0.27	0.33	0.07
	(0.06, 0.55)	(0.13, 0.61)	(-0.00, 0.26)	(0.05, 0.50)	(0.12, 0.56)	(-0.00, 0.24)

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 macro variables GRO and INF ($\bar{R}^2_2$), as well as the difference in adjusted $\bar{R}^2$. GRO is the three-month moving average of the Chicago Fed National Activity Index, and INF is one-year expected inflation measured by Blue Chip inflation forecasts. The data used for the left half of the table is the original data set of Joslin et al. (2014); the data used in the right half is extended to December 2013. The last panel shows results for the average excess bond return for all bond maturities from two to ten years. The first row of each panel reports the values of the statistics in the original data. The next three rows report bootstrap small-sample mean, and the 95%-confidence intervals (in parentheses). The bootstrap simulations are obtained under the null hypothesis that the macro variables have no predictive power. The bootstrap procedure for the simple bootstrap and the bias-corrected (BC) bootstrap is described in the main text.

Table 5: Joslin-Priebsch-Singleton: inference in excess return regressions

	$PC1$	$PC2$	$PC3$	GRO	INF	Wald
Original sample: 1985–2008						
Coefficient	1.064	1.988	3.342	-2.174	-6.494	
HAC statistic	5.603	4.671	0.865	2.438	4.232	25.476
HAC p-value	**0.000**	**0.000**	0.388	**0.015**	**0.000**	**0.000**
Bootstrap 5% c.v.				3.203	3.950	24.410
Bootstrap p-value				0.129	**0.038**	**0.046**
BC bootstrap 5% c.v.				3.460	4.286	27.664
BC bootstrap p-value				0.140	0.052	0.061
IM $q = 8$	**0.002**	**0.040**	**0.002**	0.563	0.940	
IM $q = 16$	**0.003**	**0.002**	0.063	0.244	0.500	
Estimated size of tests						
HAC				0.209	0.285	0.382
Simple bootstrap				0.058	0.067	0.069
IM $q = 8$				0.049	0.054	
IM $q = 16$				0.038	0.033	
Later sample: 1985–2013						
Coefficient	0.523	1.865	4.330	-0.271	-3.767	
HAC statistic	2.524	3.755	1.345	0.323	2.408	5.799
HAC p-value	**0.012**	**0.000**	0.180	0.747	**0.017**	0.055
Bootstrap 5% c.v.				3.332	3.665	22.786
Bootstrap p-value				0.820	0.178	0.376
BC bootstrap 5% c.v.				3.420	3.919	24.471
BC bootstrap p-value				0.838	0.206	0.417
IM $q = 8$	0.275	**0.030**	**0.003**	0.550	0.325	
IM $q = 16$	0.304	**0.007**	0.139	0.393	0.934	

Predictive regressions for annual excess bond returns, averaged over two- through ten-year bond maturities, using yield PCs and macro variables (which are described in the notes to Table 4). The data used for the top panel is the original data set of Joslin et al. (2014); the data used for the bottom panel is extended to December 2013. 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. 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. See the text for a description of the experimental design for the simple bootstrap and the bias-corrected (BC) bootstrap. We also report p-values for t-tests using the methodology of Ibragimov and Müller (2010) (IM), splitting the sample into either 8 or 16 blocks. The last four rows in the first panel report bootstrap estimates of the true size of different tests with 5% nominal coverage, calculated as the frequency of bootstrap replications in which the test statistics exceed their critical values, except for the size of bootstrap test which is calculated as described in the main text. p-values below 5% are emphasized with bold face.

Table 6: Ludvigson-Ng: predicting excess returns using PCs and macro factors

	$PC1$	$PC2$	$PC3$	$F1$	$F2$	$F3$	$F4$	$F5$	$F6$	$F7$	$F8$	Wald
A. Original sample: 1964–2007												
Coefficient	0.136	2.052	-5.014	0.742	0.146	-0.072	-0.528	-0.321	-0.576	-0.401	0.551	
HAC statistic	1.552	2.595	2.724	1.855	0.379	0.608	1.912	1.307	2.220	2.361	3.036	42.084
HAC p-value	0.121	**0.010**	**0.007**	0.064	0.705	0.543	0.056	0.192	**0.027**	**0.019**	**0.003**	**0.000**
Bootstrap 5% c.v.				2.572	2.580	2.241	2.513	2.497	2.622	2.446	2.242	29.686
Bootstrap p-value				0.140	0.761	0.594	0.128	0.301	0.092	0.057	**0.010**	**0.009**
IM $q = 8$	**0.001**	**0.001**	0.225	0.098	0.558	0.579	0.088	0.703	0.496	0.085	0.324	
IM $q = 16$	**0.000**	0.052	0.813	0.228	0.317	0.771	0.327	0.358	0.209	**0.027**	0.502	
Estimated size of tests												
HAC				0.131	0.132	0.097	0.124	0.126	0.134	0.113	0.086	0.335
Bootstrap				0.058	0.055	0.053	0.061	0.055	0.053	0.049	0.046	0.061
IM $q = 8$				0.051	0.050	0.051	0.049	0.049	0.052	0.050	0.042	
IM $q = 16$				0.051	0.048	0.051	0.050	0.051	0.045	0.055	0.046	
B. Later sample: 1985–2013												
Coefficient	0.157	1.182	2.725	0.651	-0.274	0.147	-0.488	0.022	0.334	0.035	-0.075	
HAC statistic	1.506	1.111	0.682	1.652	0.267	0.690	1.162	0.038	1.866	0.153	0.423	13.766
HAC p-value	0.133	0.268	0.496	0.099	0.789	0.491	0.246	0.969	0.063	0.878	0.673	0.088
Bootstrap 5% c.v.				2.817	2.908	2.516	2.667	2.798	2.468	2.365	2.298	37.267
Bootstrap p-value				0.224	0.844	0.587	0.370	0.973	0.136	0.892	0.718	0.495
IM $q = 8$	**0.014**	**0.005**	0.068	0.139	0.511	0.537	0.899	0.767	0.144	0.923	0.398	
IM $q = 16$	**0.024**	0.185	0.788	0.831	0.636	0.923	0.187	0.570	0.882	0.703	0.239	

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 2013. The bootstrap is a simple bootstrap without bias correction. For a description of the statistics in each row, see the notes to Table 5. p-values below 5% are emphasized with bold face.

Table 7: Ludvigson-Ng: $\bar{R}^2$ for predicting excess returns using PCs and macro factors

	$\bar{R}^2_1$	$\bar{R}^2_2$	$\bar{R}^2_2 - \bar{R}^2_1$
Original sample: 1964–2007			
Data	0.25	0.35	0.10
Bootstrap	0.20	0.24	0.03
	(0.05, 0.39)	(0.08, 0.42)	(-0.00, 0.11)
Later sample: 1985–2013			
Data	0.14	0.18	0.04
Bootstrap	0.26	0.29	0.03
	(0.05, 0.49)	(0.08, 0.51)	(-0.01, 0.14)

Adjusted $\bar{R}^2$ for regressions of annual excess bond returns, averaged over two- through five-year bonds, on three PCs of the yield curve ($\bar{R}^2_1$) and on three yield PCs together with eight macro factors ($\bar{R}^2_2$), as well as the difference in $\bar{R}^2$. 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 2013. For each data sample we report the values of the statistics in the data, and the mean and 95%-confidence intervals (in parentheses) of the bootstrap small-sample distributions of these statistics. The bootstrap simulations are obtained under the null hypothesis that the macro variables have no predictive power. The bootstrap procedure, which does not include bias correction, is described in the main text.

Table 8: Ludvigson-Ng: predicting excess returns using 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.		3.809		3.799		3.874		3.898
Bootstrap p-value		**0.022**		**0.015**		**0.017**		**0.014**
Estimated size of tests								
HAC		0.514		0.538		0.545		0.539
Bootstrap		0.047		0.055		0.057		0.050
Later sample: 1985–2013								
Coefficient	0.349	0.371	0.661	0.695	1.101	0.895	1.320	1.021
HAC t-statistic	2.644	3.348	2.527	3.409	3.007	3.340	2.946	3.270
HAC p-value	**0.009**	**0.001**	**0.012**	**0.001**	**0.003**	**0.001**	**0.003**	**0.001**
Bootstrap 5% c.v.		3.890		4.014		4.026		3.942
Bootstrap p-value		0.103		0.116		0.124		0.128

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 the original data set used by Ludvigson and Ng (2010); the second panel uses a data sample that starts in 1985 and ends in 2013. HAC t-statistics and p-values are calculated using Newey-West standard errors with 18 lags. We obtain bootstrap distributions of the t-statistics under the null hypothesis that macro factors and hence $H8$ have no predictive power. We also report bootstrap critical values (c.v.'s) and p-values, as well as estimates of the true size of conventional t-tests and the bootstrap tests with 5% nominal coverage (see notes to Table 5). The bootstrap procedure, which does not include bias correction, is described in the main text. p-values below 5% are emphasized with bold face.

Table 9: Ludvigson-Ng: $\bar{R}^2$ for predicting excess returns using return-forecasting factors

	Original sample: 1985–2008			Later sample: 1985–2013		
	$\bar{R}^2_1$	$\bar{R}^2_2$	$\bar{R}^2_2 - \bar{R}^2_1$	$\bar{R}^2_1$	$\bar{R}^2_2$	$\bar{R}^2_2 - \bar{R}^2_1$
Two-year bond						
Data	0.31	0.42	0.11	0.15	0.23	0.07
Bootstrap	0.21	0.24	0.03	0.25	0.28	0.03
	(0.06, 0.39)	(0.09, 0.41)	(-0.00, 0.10)	(0.04, 0.50)	(0.08, 0.52)	(-0.00, 0.12)
Three-year bond						
Data	0.33	0.43	0.10	0.15	0.22	0.07
Bootstrap	0.20	0.23	0.03	0.25	0.29	0.04
	(0.05, 0.38)	(0.09, 0.40)	(-0.00, 0.10)	(0.05, 0.48)	(0.09, 0.51)	(-0.00, 0.13)
Four-year bond						
Data	0.36	0.45	0.09	0.19	0.24	0.05
Bootstrap	0.21	0.25	0.03	0.27	0.30	0.03
	(0.06, 0.40)	(0.10, 0.42)	(-0.00, 0.11)	(0.07, 0.50)	(0.11, 0.52)	(-0.00, 0.12)
Five-year bond						
Data	0.33	0.42	0.09	0.17	0.21	0.05
Bootstrap	0.21	0.24	0.03	0.25	0.29	0.03
	(0.06, 0.39)	(0.10, 0.41)	(-0.00, 0.11)	(0.06, 0.48)	(0.10, 0.50)	(-0.00, 0.13)

Adjusted $\bar{R}^2$ for regressions of annual excess bond returns on return-forecasting factors based on yield-curve information (CP) and macro information ($H8$), as in Ludvigson and Ng (2010). $\bar{R}^2_1$ is for regressions with only CP, while $\bar{R}^2_2$ is for regressions with both CP and $H8$. The table shows results both for the original data set used by Ludvigson and Ng (2010) and for a data sample that starts in 1985 and ends in 2013. For each data sample and bond maturity, we report the values of the statistics in the data, and for the bootstrap small-sample distributions of these statistics the mean, and 95%-confidence intervals (in parentheses). The bootstrap simulations are obtained under the null hypothesis that the macro variables have no predictive power. The bootstrap procedure, which does not include bias correction, is described in the main text.

Table 10: Cochrane-Piazzesi: in-sample evidence

	$PC1$	$PC2$	$PC3$	$PC4$	$PC5$	Wald	R_1^2	R_2^2	$R_2^2 - R_1^2$
Original sample: 1964–2003									
Data	0.127	2.740	-6.307	-16.128	-2.038		0.26	0.35	0.09
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./mean $\bar{R}^2$				2.253	2.236	8.464	0.21	0.21	0.01
Bootstrap p-value/95% CIs				**0.000**	0.507	**0.000**	(0.05, 0.40)	(0.06, 0.41)	(0.00, 0.03)
IM $q = 8$	**0.002**	**0.030**	0.873	0.237	0.233				
IM $q = 16$	**0.000**	**0.004**	0.148	0.953	0.283				
Estimated size of tests									
HAC				0.085	0.083	0.114			
Bootstrap				0.046	0.053	0.055			
IM $q = 8$				0.040	0.050				
IM $q = 16$				0.043	0.049				
Later sample: 1985–2013									
Data	0.104	1.586	3.962	-9.196	-9.983		0.14	0.17	0.03
HAC statistic	1.619	2.215	1.073	1.275	1.351	4.174			
HAC p-value	0.106	**0.027**	0.284	0.203	0.178	0.124			
Bootstrap 5% c.v./mean $\bar{R}^2$				2.463	2.433	9.878	0.26	0.28	0.02
Bootstrap p-value/95% CIs				0.301	0.273	0.272	(0.06, 0.49)	(0.08, 0.50)	(0.00, 0.05)
IM $q = 8$	**0.011**	0.079	**0.044**	0.803	0.435				
IM $q = 16$	**0.001**	**0.031**	0.215	0.190	0.949				

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. We also report the unadjusted R^2 for the regression using only three PCs (R_1^2) and for the regression including all five PCs (R_2^2), as well as the difference in these two. Bootstrap distributions are obtained under the null hypothesis, using the bootstrap procedure described in the main text (without bias correction). For the R^2-statistics, we report means and 95%-confidence intervals (in parentheses). For the HAC test statistics, bootstrap 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 the statistic in the data. We also report p-values for t-tests using the methodology of Ibragimov and Müller (2010) (IM), splitting the sample into either 8 or 16 blocks. The last four rows in the first panel report bootstrap estimates of the true size of different tests with 5% nominal coverage, calculated as the frequency of bootstrap replications in which the test statistics exceed their critical values, except for the size of bootstrap test which is calculated as described in the main text. p-values below 5% are emphasized with bold face.

Table 11: Cochrane-Piazzesi: out-of-sample forecast accuracy

n	R_2^2	R_1^2	$RMSE_2$	$RMSE_1$	DM	p-value	$RMSE_{mean}$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
2	0.321	0.260	2.120	1.769	2.149	0.034	1.067
3	0.341	0.242	4.102	3.232	2.167	0.032	1.946
4	0.371	0.266	5.848	4.684	2.091	0.039	2.989
5	0.346	0.270	7.374	6.075	2.121	0.036	3.987
average	0.351	0.264	4.845	3.917	2.133	0.035	2.385

In-sample vs. out-of-sample predictive power for excess bond returns (averaged across maturities) of a restricted model with three PCs and an unrestricted model with five PCs. The in-sample period is from 1964 to 2002 (the last observation used by Cochrane-Piazzesi), and the out-of-sample period is from 2003 to 2013. The second and third column show in-sample R^2. The fourth and fifth column show root-mean-squared forecast errors (RMSEs) of the two models. The column labeled "DM" reports the z-statistic of the Diebold-Mariano test for equal forecast accuracy, and the following column the corresponding p-value. The last column shows the RMSE when forecasts are the in-sample mean excess return.

Table 12: 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**
IM $q = 8$	**0.030**					0.795
IM $q = 16$	**0.001**					0.925
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
IM $q = 8$	**0.006**	**0.013**				0.972
IM $q = 16$	**0.000**	**0.000**				0.557
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
IM $q = 8$			**0.000**	**0.003**	**0.045**	0.968
IM $q = 16$			**0.000**	**0.000**	**0.023**	0.854
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.						3.105
Bootstrap p-value						0.448
IM q = 8			**0.005**	0.134	0.714	0.494
IM q = 16			**0.008**	**0.011**	0.611	0.980

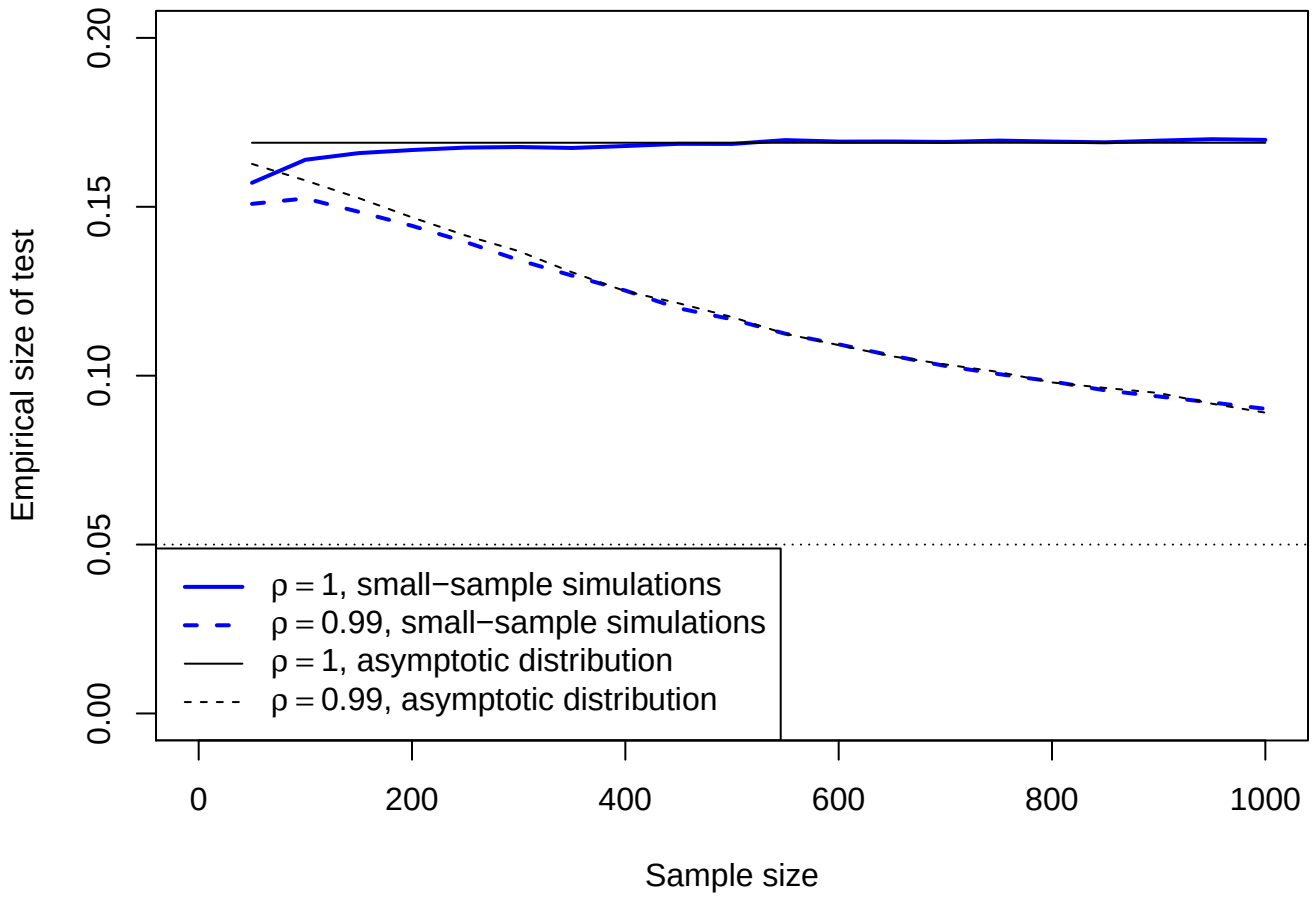
Predictive regressions for annual bond returns using Treasury bond supply, as in Greenwood and Vayanos (2014) (GV). The coefficients on bond supply in the first two panels are identical to those reported in row (1) and (6) of Table 5 in GV. HAC t-statistics and p-values are constructed using Newey-West standard errors with 36 lags, as in GV. The last 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 13: Cooper-Priestley: predictive power of the output gap

	gap	$\tilde{CP}$	CP	$PC1$	$PC2$	$PC3$
Coefficient	-0.126					
OLS t-statistic	3.224					
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	0.272		**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			1.354	4.362	2.507
HAC p-value	0.192			0.176	**0.000**	0.012
Bootstrap 5% c.v.	2.933					
Bootstrap p-value	0.356					
IM $q = 8$	0.612			**0.002**	**0.011**	0.234
IM $q = 16$	0.243			**0.000**	**0.001**	0.064

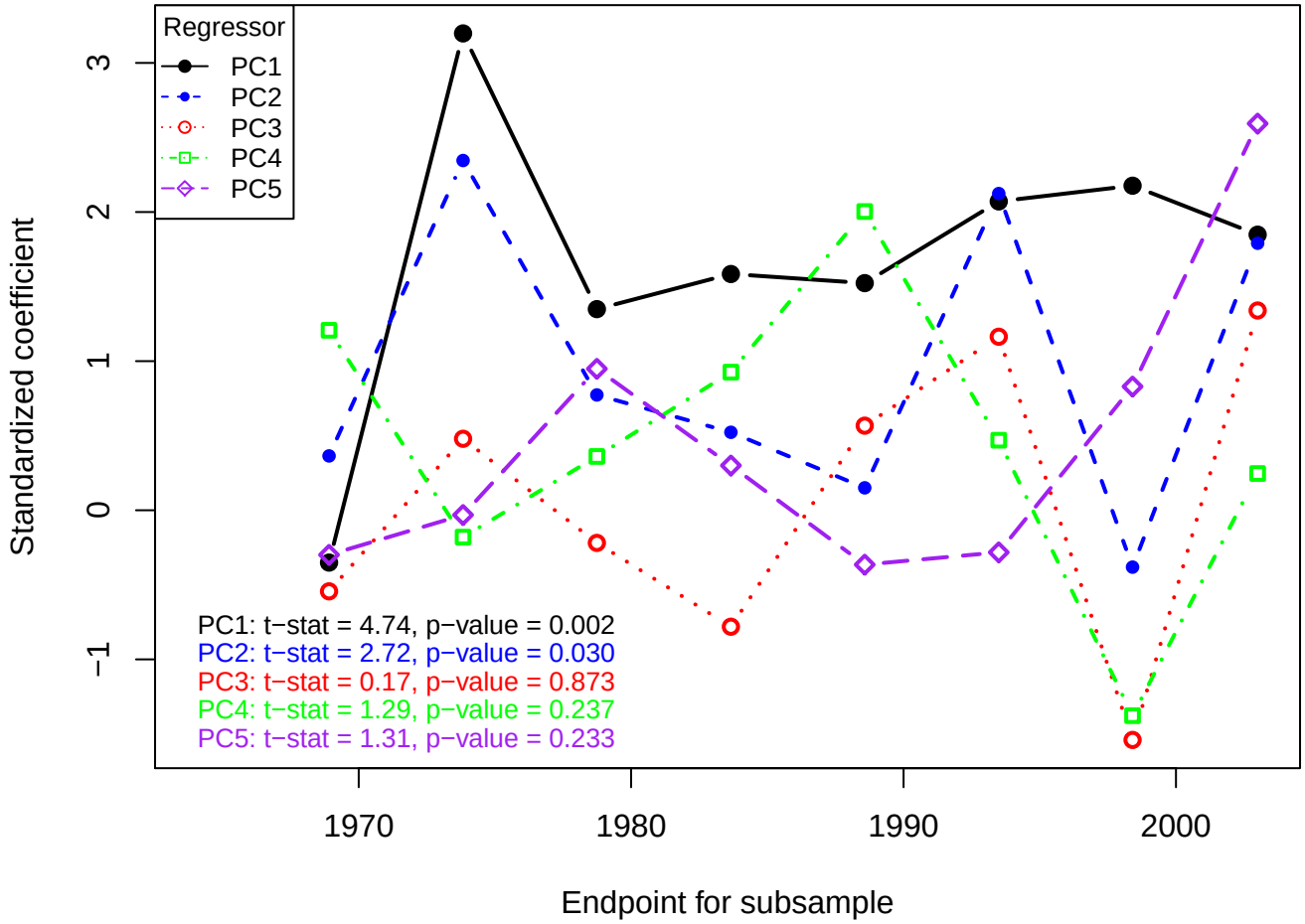
Predictive regressions for the one-year excess return on a five-year bond using the output gap, as in Cooper and Priestley (2008) (CPR). $\tilde{CP}$ is the Cochrane-Piazzesi factor after orthogonalizing it with respect to *gap*, whereas CP is the usual Cochrane-Piazzesi factor. For the predictive regression, *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.

Figure 1: Simulation study: size of t-test and sample size



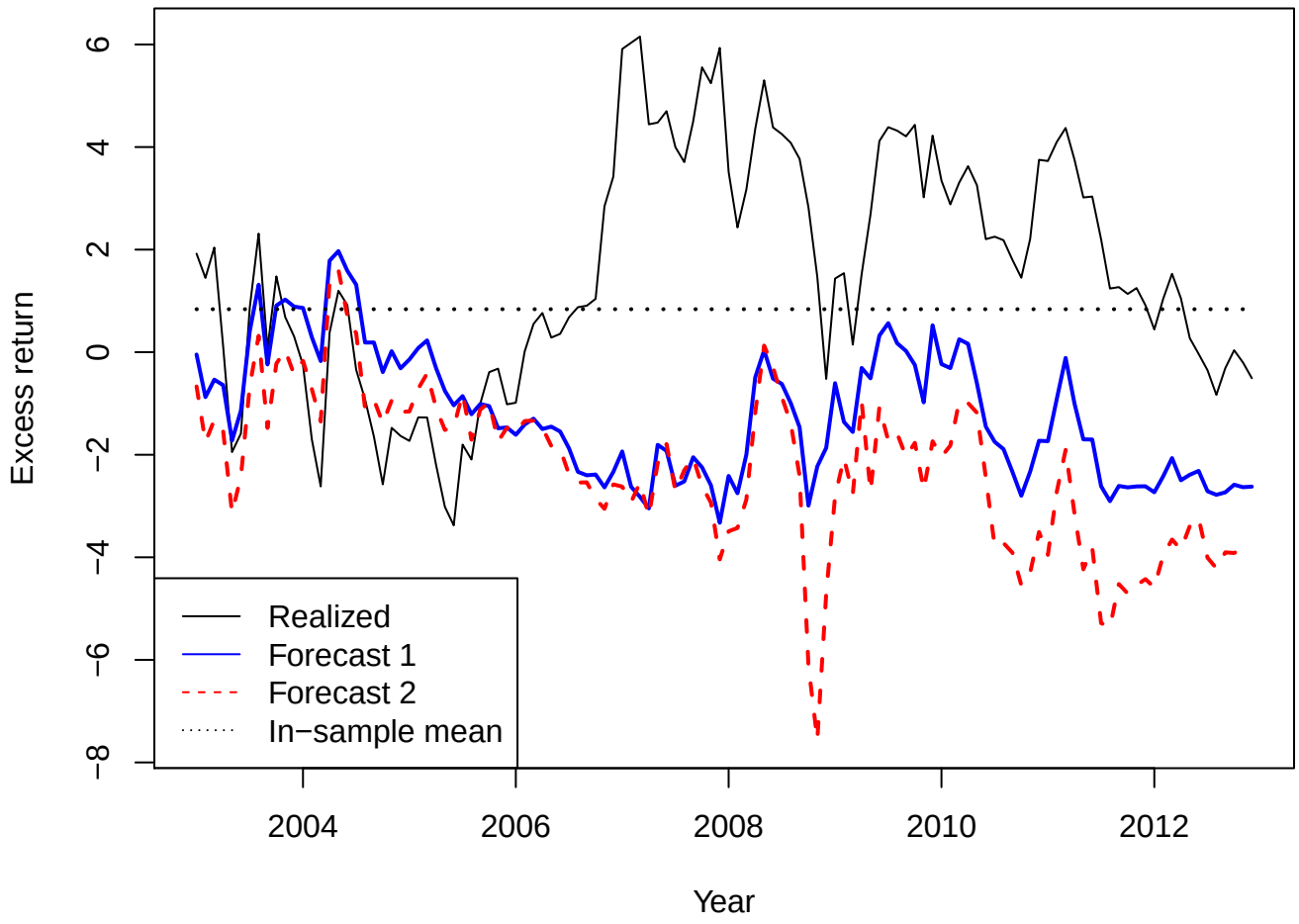
True size of conventional t-test of $H_0 : \beta_2 = 0$ with nominal size of 5%, in simulated small samples and according to local-to-unity asymptotic distribution, for different sample sizes, with $\delta = 1$. Regressors are either random walks ($\rho = 1$) or stationary but highly persistent AR(1) processes ($\rho = 0.99$). For details on the simulation study refer to main text.

Figure 2: Cochrane-Piazzesi: predictive power of PCs across subsamples



Standardized coefficients on principal components (PCs) across eight different subsamples, ending at the indicated point in time. Standardized coefficients are calculated by dividing through the sample standard deviation of the coefficient across the eight samples. Text labels indicate t-statistics and p-values of the Ibragimov-Mueller test with $q = 8$. Note that the t-statistics are equal to means of the standardized coefficients multiplied by $\sqrt{8}$. The data and sample period is the same as in Cochrane and Piazzesi (2005).

Figure 3: Cochrane-Piazzesi: out-of-sample forecasts



Realizations vs. out-of-sample forecasts of excess bond returns (averaged across maturities) from restricted model (1) with three PCs and unrestricted model (2) with five PCs. The in-sample period is from 1964 to 2002 (the last observation used by Cochrane-Piazzesi), and the out-of-sample period is from 2003 to 2013. The figure also shows the in-sample mean excess return.