Machine-Learning-Based Return Predictors and the Spanning Controversy in Macro-Finance*

Jing-Zhi Huang†
Penn State University

Zhan Shi‡
Tsinghua University

September 15, 2021

Abstract

We propose a two-step machine learning algorithm—the Supervised Adaptive Group LASSO (SAGLasso) method—that is suitable for constructing parsimonious return predictors from a large set of macro variables. We apply this method to government bonds and a set of 917 macro variables and construct a new, transparent, and easy-to-interpret macro variable with significant out-of-sample predictive power for excess bond returns. This new macro factor, termed the SAGLasso factor, is a *linear* combination of merely 30 selected macro variables out of 917. Furthermore, it can be decomposed into three sub-level factors: a novel "housing" factor, an "employment" factor, and an "inflation" factor. Importantly, the predictive power of the SAGLasso factor is robust to bond yields; namely, the SAGLasso factor is not spanned by bond yields. Moreover, we show that the unspanned variation of the SAGLasso factor cannot be attributed to yield measurement error or macro measurement error. The SAGLasso factor therefore provides a potential resolution to the spanning controversy in the macro-finance literature.

***Management Science*, forthcoming**

*We are very grateful to Haoxiang Zhu (Finance Department Editor), an Associate Editor, and an anonymous referee for their extensive and detailed comments that helped to improve the article substantially. We also thank Yakov Amihud, Gurdip Bakshi, Charles Cao, Long Chen, Mike Chernov, Ed Coulson, Heber Farnsworth, Peter Feldhütter (AFA discussant), Laura Field, Itay Goldstein, Olesya Grishchenko, Raymond Kan, Anh Le, Hong Liu, Stefan Nagel, Matt Pritsker (FIC discussant), Marco Rossi, Ilona Shiller (FMA discussant), Andrea Tamoni, Dan Thornton, Joel Vanden, Hong Yan (SIF discussant), Weina Zhang (CICF discussant), Wei Zhong, and Hao Zhou; seminar participants at Fordham, National University of Singapore, Penn State Smeal College, Penn State Mathematics Department, Federal Reserve Bank of Philadelphia, Singapore Management University, Federal Reserve Bank of St. Louis, Temple University, University of Rhode Island, University of Waterloo, and University of Wisconsin-Milwaukee; and participants at the 20th FDIC Derivatives Securities and Risk Management Conference, the 2010 CICF, the 2010 FMA, the 2010 Summer Institute of Finance Conference, the 2011 AFA, and the 2012 Darla Moore Fixed Income Conference for valuable comments and suggestions. This paper is a new incarnation of the previous work circulated under the title "Determinants of Bond Risk Premia". We acknowledge a grant from the Penn State Institute for Real Estate Studies for partial support. We also thank Aaron Henrichsen and Terry O'Brien for editorial assistance.

†Smeal College of Business, Penn State University, University Park, PA 16802, USA. Email: jxh56@psu.edu.

‡PBC School of Finance, Tsinghua University, Beijing, 100083, China. Email: shizh@pbcsf.tsinghua.edu.cn.

1 Introduction

A growing literature has documented that excess returns of U.S. Treasury bonds are predictable. For instance, the predictors found thus far include forward rates (Cochrane and Piazzesi 2005) and yield-based variables constructed by using filtering (Duffee 2011),[1] as well as macroeconomic variables (e.g., Cooper and Priestley 2009; Ludvigson and Ng 2009). One debate in this literature is whether macroeconomic fundamentals have any such predictive power conditionally over bond yields. Among other things, this debate has important implications for macro-finance term structure models (MTSMs; see, e.g., Joslin, Priebsch, and Singleton 2014 (hereafter JPS)).

In this paper, we construct a new macro factor with strong and robust predictive power for bond risk premia using a two-step machine learning algorithm, termed the Supervised Adaptive Group LASSO (SAGLasso) method. We obtain the new macro variable (referred to as the SAGLasso factor) by applying the SAGLasso algorithm to a panel of 131 macro variables (along with six of their lags)—a total of 917 (131×7) macro variables—that are adjusted for data revisions and publication lags. In addition to its predictive power, the SAGLasso factor has two other noteworthy features. One is that the factor is parsimonious, transparent, and easy to interpret. The SAGLasso factor is a *linear* combination of merely 30 selected variables out of 917. Furthermore, it can be decomposed into three sub-level factors: a novel "housing" factor, an "employment" factor, and an "inflation" factor—which consist of 13, 11, and 6 macro variables, respectively. The other feature is that the SAGLasso factor is unspanned. Intuitively, this means that the SAGLasso factor is not subsumed (spanned) by yield factors in either predictive regressions or MTSMs. As such, the SAGLasso factor can potentially help resolve the spanning controversy in the macro-finance literature—the debate on whether macro-based return predictors are spanned or not.

We begin our analysis by describing the two-step SAGLasso method, followed by its implementation using the panel of 131 macro series. We construct eight sub-level factors—such as the

[1] See also Fama and Bliss (1987), Stambaugh (1988), and Campbell and Shiller (1991).

"housing," "employment," and "inflation" factors—in the first step and then the SAGLasso factor in the second step of the procedure. Note that we control for contemporaneous yields in both steps to minimize the information overlap between the SAGLasso factor and the yield curve.

Next, we examine the conditional predictive power of the SAGLasso factor for bond risk premia by testing two hypotheses. The first one, Spanning Hypothesis I, states that macro variables have no incremental predictive power over the current yield curve, the first three principal components (PCs) of yields. The second one, Spanning Hypothesis II—a stronger version of the first hypothesis—posits that macro variables have no incremental predictive power over the filtration generated by the yield curve, proxied by the first five yield PCs filtered from a dynamic term structure model. Our results from both in-sample and out-of-sample tests strongly reject the two spanning hypotheses when the SAGLasso factor is the sole macro variable used. These results indicate that the SAGLasso macro factor has significant incremental predictive power, over price-related information in the Treasury market, for future bond returns. Furthermore, we provide evidence that this predictability can generate significant economic gains for investors.

Lastly, as an important application of the SAGLasso factor, we revisit the spanning controversy. Given that the SAGLasso factor has strong predictive power for bond risk premia yet is weakly correlated with the current yield curve, the new macro factor may shed light on the controversy. To this end, we examine three aspects of the controversy using the JPS framework for MTSMs. First, we show that the conditional predictive power of the SAGLasso factor is robust to finite sample tests. Second, we focus on part of the spanning controversy formulated under the MTSM framework and test the macro-unspanning hypothesis (MUH), which says that a given MTSM's macro state variables are not spanned by its yield factors.[2] We find that when an $\mathcal{N}$-factor MTSM with $4 \leq \mathcal{N} \leq 6$ includes the SAGLasso factor as its sole macro factor, our likelihood ratio tests do not reject the MUH, thereby presenting statistical evidence on the relevance of unspanned MTSMs. Third, we provide confirmative evidence that the temporal variation in the SAGLasso factor is not

[2]Such models are referred to as unspanned MTSMs. Models with spanned macro risks are called spanned models.

spannned/explained by the current yield curve. Importantly, this result is robust to measurement errors in yields or in the SAGLasso macro variable itself. Taken together, these findings suggest that the SAGLasso factor provides a potential resolution to the spanning controversy.

To summarize, this study contributes to the macro finance literature in three dimensions. First, it is among the first to introduce a machine learning algorithm suitable for constructing parsimonious return predictors from a large set of macro variables. Second, using this algorithm we construct a new, easy-to-interpret macro variable that has strong out-of-sample conditional predictive power for bond risk premia. Moreover, unlike commonly used macro variables in the literature, the SAGLasso factor is unspanned and has tiny measurement error. Third, we show that, due to its unique features, the SAGLasso factor can address those concerns raised in Bauer and Rudebusch (2016; hereafter BR), Bauer and Hamilton (2018; hereafter BH), and Ghysels, Horan, and Moench (2018) in a unified manner and thus can potentially help resolve the spanning controversy.

While this paper focuses on linear models of predictors, two related studies use nonlinear machine learning models to construct bond return predictors (but do not address the spanning controversy). Huang et al. (2016) find that the macro series selected by SAGLasso is robust to various nonlinear models they consider. Bianchi, Büchner, and Tamoni (2021) study bond risk premia using tree-based algorithms as well as neural networks and find that their superb statistical performance translates into large economic gains. While these highly nonlinear methods can accommodate more complex data, the SAGLasso method can lead to easier-to-interpret return predictors.[3]

The remainder of the paper is organized as follows. Section 2 states Spanning Hypotheses I & II, followed by Section 3 on the data we use. Section 4 presents the SAGLasso algorithm, constructs the SAGLasso factor, and examines its properties. Section 5 revisits the spanning controversy. Section 6 concludes. Appendix A lists some notation and terms frequently used in the paper.

[3]Several other studies focus on the application of machine learning in the other finance markets. Freyberger et al. (2020) use Group Lasso to study the impact of characteristics on expected stock returns. Gu et al. (2020) compare Group Lasso with other machine learning methods in the context of stock return prediction. Bali et al. (2021) and He et al. (2021) apply nonlinear machine learning models to inferring corporate bond risk premiums.

2 Hypotheses on the Predictive Power of Macro Variables

2.1 Basic Setup

We use continuously compounded annual log returns on an n-year zero-coupon Treasury bond in excess of the annualized yield on a one-year zero-coupon Treasury bond. That is, for $t = 1, \cdots, T$, excess returns $rx_{t,t+12}^{(12n)} = r_{t,t+12}^{(12n)} - y_t^{(12)} = ny_t^{(12n)} - (n-1)y_{t+12}^{(12(n-1))} - y_t^{(12)}$, where $r_{t,t+12}^{(12n)}$ is the one-year log holding-period return on an n-year bond purchased at the end of month t and sold at the end of month $t+12$, and $y_t^{(12n)}$ is the time-t log yield on the n-year bond.

We consider the following predictive regression that is often used to investigate the role of the macroeconomy in shaping bond risk premia (e.g., Ludvigson and Ng 2009 and JPS):

$$rx_{t,t+12}^{(12n)} = \alpha + \beta_Z' Z_t + \beta_F' F_t + e_{t+12}, \tag{1}$$

where Z represents yield curve factors that are supposed to summarize yield-based information in the Treasury bond market and F denotes macroeconomic factors. For example, Z can be factors constructed from the current yield curve (e.g., yield spreads used in Campbell and Shiller 1991 or return predictors estimated using historical yields (e.g., the Cochrane-Piazzesi forward rate factor). Similarly, F can be either predetermined macroeconomic measures (e.g., the GDP growth and NAPM price index) or factors extracted from a set of macroeconomic series, such as the Ludvigson and Ng (2009; LN09 hereafter) factor and the new macro factors constructed in this study. The remainder of this section focuses on null hypotheses about the predictive power of macro variables and whether they are spannned.

2.2 Spanning Hypotheses

The issue of interest is macro factors' conditional predictive power above and beyond that contained in the yield curve. Empirically, this issue can be examined based on the significance of β_F in Eq. (1), for a given Z_t.

It is known that the first three principal components (PCs) of yields explain all but a negligible

fraction of the variation in the term structure (Litterman and Scheinkman 1991). If the current yield curve is supposed to contain almost all the information useful for determining term premia, we arrive at Spanning Hypothesis I (a hypothesis formulated and tested by JPS and BH):

$$H_0^{S1} : \text{in Eq. (1), if } Z_t = PC^o_{1-3,t}, \text{ then } \beta_F = 0,$$

where $PC^o_{1-3} = (PC^o_1, PC^o_2, PC^o_3)$, the vector of the first three PCs of the *observed* yield curve.

Interestingly, Duffee (2011) finds that the fourth and fifth PCs are also informative about predicting bond returns. These factors need to be estimated using filtering techniques based on both current and historical yields, however, as the effects of such factors on cross-sectional yields are too small to dominate measurement error in observed yields. Nonetheless, a natural question is whether macro variables contain information about future bond returns that is not captured by the filtration generated by the yield curve process. If the "true" yield curve is Markov, as is commonly assumed in term structuring modeling, this question leads to Spanning Hypothesis II:

$$H_0^{S2} : \text{in Eq. (1), if } Z_t = PC_{1-5,t}, \text{ then } \beta_F = 0,$$

where $PC_{1-5} = (PC_1, \ldots, PC_5)$, the vector of the first five PCs of the *noise-uncontaminated* yield curve. Given the predictive power of filtered PC_{4-5}, H_0^{S2} provides a stronger test of the conditional predictive power of F_t than does H_0^{S1}.[4] We also consider an alternative version of H_0^{S2} where Z_t is the spanned "cycle" factor of Cieslak and Povala (2015) in Internet Appendix IA.F.

Small-sample distortions may also take place in tests of H_0^{S1} and H_0^{S2}. BH demonstrate that estimates of standard errors in the t-test of $\beta_F = 0$ can be biased because PCs (Z_t) are typically persistent and autoregressive with innovation terms that are possibly correlated with e_{t+12}. They propose a bootstrap procedure to account for the size distortion and conclude that much of extant "evidence against the spanning hypothesis is in fact spurious." Besides the statistical inference about β_F in Eq. (1), BH also study the finite-sample distribution of the increase in R^2 when F_t is added to the regression. They show that serially correlated e_{t+12} due to overlapping observations could

[4]The use of PC_{4-5} rather than PC^o_{4-5} in H_0^{S2} is because the latter's predictive power is weaker (see Internet Appendix IA.A). The version of H_0^{S2} based on PC^o_{4-5} is examined in JPS, BR, and BH.

substantially inflate the incremental R^2 in small samples, even if F_t provides no help in predicting bond returns. We test H_0^{S1} and H_0^{S2} by conducting an asymptotic inference (Sections 4.4.2) as well as an MTSM-based finite-sample inference (Section 5.2).

3 Data

We use monthly data on bond returns and macroeconomic variables over the period January 1964 to December 2014 in our analysis. The start of our sample coincides with that of many other studies that also use the Fama-Bliss yield data set (e.g., Cochrane and Piazzesi 2005; Ludvigson and Ng 2009). We also conduct part of the empirical analysis based on the 1985–2014 subsample because, first, several studies including JPS and BR focus on post-1984 samples; secondly, some studies argue that the predictive power of macro variables weakens in more recent samples, especially post-1984;[5] and thirdly, the vintage data coverage for many time series starts in the early 1980s.

Bond data used in this study consist of monthly prices for one- through five-year zero-coupon Treasury bonds from the CRSP (Fama Risk Free Rates and Fama-Bliss Discount Bond Yields) for the full sample, and self-constructed monthly zero yields with maturity beyond five but up to ten years for the post-1984 sample. The latter data set extends the original Fama-Bliss data to longer maturities and is constructed using monthly quotes on individual bonds from the CRSP Master File of Treasury Bonds by following Le and Singleton (2013).[6] Zero yields can then be used to construct annual excess returns as defined in Section 2.1.

Our macro data set consists of a balanced panel of 131 monthly macroeconomic times series, and is an updated and "real-time" version of the macro data set used in Stock and Watson (2002, 2005) and LN09 that includes one more economic series no longer available. The main source of

[5]For instance, BH find that the predictive power of macro variables is substantially weaker in extended samples that include observations in 2010s; BH also question the stability of Ludvigson and Ng's results for their macro return predictors across different subsample periods, especially over the post-1984 sample. Additionally, Duffee (2013a,b) notes that "the predictability associated with Ludvigson and Ng's real activity factor may be sample-specific." Our main results are also robust to a backward sample extension to 1952, the starting year of the original Fama-Bliss data (Internet Appendix IA.B).

[6]Similarly extended Fama-Bliss data are used in JPS and BR. An alternative data set used in the literature is constructed by Gürkaynak, Sack, and Wright (2007).

our real-time macro data is the Archival Federal Reserve Economic Data (ALFRED) database at the Federal Reserve Bank of St. Louis, which is a collection of vintage versions of U.S. economic data and contains more monthly sampled series than does the Philadelphia Fed's Real-Time Data Set. Appendix B includes the list of the 131 series in Table A.1 and describes how our macro data are compiled. The 131 series are organized in a hierarchical manner. Such a cluster structure of macro variables turns out to be useful to model selection. To that end, following Ludvigson and Ng (2011), we group the 131 series into eight categories: i) output (17 series); ii) labor market (32 series); iii) housing sector (10 series); iv) orders and inventories (14 series); v) money and credit (11 series); vi) bond and FX—interest rates or financial (22 series); vii) prices or price indices (21 series); and viii) stock market (4 series). Column 2 of Table A.1 reports the group ID of each series. Section 4.2 shows that some of the eight groups have stronger predictive power than the others.

4 Adaptive-Lasso-Based Model Selection

In this section we first describe the supervised adaptive group lasso algorithm. We next use the algorithm to construct a macro factor with low correlations to the yield curve. We then examine the predictive power of this new macro factor for future bond returns as well as economic gains of such bond return predictability.

4.1 Supervised Adaptive Group Lasso

For a $T \times 1$ response vector $\mathbf{y}$, consider the following penalized least squares (PLS) function:

$$f^{\text{PLS}}(\beta) = \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda \sum_{i=1}^{N} |\beta_i|, \tag{2}$$

where $\lambda \geq 0$ is a tuning parameter used to penalize the complexity of the model, and $\|\cdot\|$ is the ℓ_2-norm, namely, $\|\eta\| := (\eta'\eta)^{1/2}, \forall \eta \in \mathbb{R}^T$. The ℓ_1-norm penalty $|\beta_i|$ used here induces sparsity in the solution and defines the "least absolute shrinkage and selection operator" (Tibshirani 1996)—this method is usually referred to as "lasso" rather than "LASSO" in the statistics literature. The

lasso estimate is given by $\hat{\beta}^{lasso} = \arg\min_{\beta} f^{\text{PLS}}(\beta)$.

If λ is zero, then $\hat{\beta}^{lasso}$ equals the OLS estimate, $\hat{\beta}^{ols}$, provided that the OLS estimation is feasible. Recall that none of $\hat{\beta}^{ols}$'s components are zero. However, as λ increases, some components of $\hat{\beta}^{lasso}$ will shrink to zero, and as a result, the corresponding "useless" explanatory variables will be dropped and the resulting regression model will become more parsimonious.

Lasso has several advantages over the OLS. First, by construction, lasso reduces the variance of the predicted value and thus improves the overall (out-of-sample) forecasting performance. Second, the OLS is known to have poor finite sample properties when the dimension of parameters to be estimated is comparable with the number of observations. For instance, in our case there are 131 macro series along with six of their lags—917 (131×7) macro variables in total—with only 600 observations for each series. Lasso is developed to handle such problems. Third, lasso leads to a much more parsimonious and easier-to-interpret model than the OLS. In fact, the parsimonious or sparse feature of lasso distinguishes it from ridge regression, another shrinkage method.

Despite lasso's popularity, one limitation of the method is that lasso estimates can be biased. Zou (2006) shows that this problem can be fixed by using Adaptive Lasso, which minimizes the following objective function:

$$\|\mathbf{y} - \mathbf{X}\beta\|^2 + \sum_{i=1}^{N} \lambda_i |\beta_i|, \tag{3}$$

where different tuning parameters $\{\lambda_i\}$ are introduced to penalize different β_is separately.

We construct a macro-based return predictor in two steps. In the first step, we utilize the cluster structure of our macroeconomic panel and consider variable selection separately within each of the eight groups/clusters formed in Section 3; that is, we screen out less important or irrelevant individual economic series and identify informative ones within each cluster using adaptive lasso. This is done for three reasons. First, even variables within the same group may represent certain quantitative measurements of different economic sectors. For instance, the Industrial Production (IP) Index of Consumer Goods and the IP Index of Materials (in group i) might be connected

to bond risk premia in a different manner. Second, we want to select macroeconomic measures that are jointly significantly associated with bond risk premia. Third, adaptive lasso selects only a small number of macro variables within each cluster and thus allows us to construct parsimonious models, including easy-to-interpret group macro factors if necessary.

In the second step, we consider all the groups together, each of which now consists of only those macro variables selected in step one, and then conduct variable selection at the group level. We implement this idea using the Group Lasso of Yuan and Lin (2006) to deal with situations in which covariates are assumed to be clustered in groups (see Appendix C). That is, we select important clusters using group lasso, thereby identifying influential economic sectors in addition to individual variables selected in the first step.[7]

We refer to this two-step procedure as the supervised adaptive group lasso (SAGLasso) algorithm.[8] Its key feature is to consider penalized time-series selection at both the within-cluster level and the cluster level. We construct bond return predictors by applying SAGLasso to a large set of macro series in this study. SAGLasso should also be useful in similar big data applications in finance and economics.

4.2 A Macro-Based Return-Forecasting Factor

This subsection implements the two-step SAGLasso procedure using the average excess return (the bond market return), $arx_{t,t+12} = \frac{1}{(n_b-1)} \sum_{n=2}^{n_b} rx_{t,t+12}^{(n)}$, as the dependent variable, where n_b equals 5 (10) when the full (post-1984) sample is used.

First, we perform model selection in each of the eight groups of macro series separately, using only macro variables within the same group along with their six lagged values. To minimize the

[7]Using high-dimension model selection (e.g., Huang, Shi, and Zhong 2015), Huang, Li, Ni, and Shi (2016) find that the variables selected under the SAGLasso procedure are robust to a variety of nonlinear models. Bianchi, Büchner, and Tamoni (2021) also emphasize that it is important to exploit the cluster structure of the macroeconomic panel and do selection within groups and across groups. As such, different machine learning methods seemingly can capture the "common" cluster structure of the same macro data, at least for the purpose of bond return predictions.

[8]In statistical learning, a problem is considered to be supervised if the goal is to predict the value of an outcome measure based on a variety of input measures. See Appendix C for more details of the SAGLasso procedure.

information overlap with respect to yield curve factors, we include the first three yield PCs in our variable selection but do not penalize the associated coefficients. Put differently, in the regression framework of Eq. (1), Z_t is $PC^o_{1-3,t}$ but β_Z are not penalized; F_t includes contemporaneous and lagged macro variables in a given group and β_F are subject to shrinkage. Therefore, at the intragroup level of group j, we minimize the following objective function:

$$\|\mathbf{arx} - \mathbf{Z}\boldsymbol{\beta}^{\langle 1\rangle}_{Z,j} - \mathbf{F}\boldsymbol{\beta}^{\langle 1\rangle}_{F,j}\|^2 + \sum_{i=1}^{7N_j} \lambda^j_i |\beta^{\langle 1\rangle}_{F,j,i}|,$$

where λ^j_i is the tuning parameter; N_j denotes the number of economic series contained in group j; $\beta^{\langle 1\rangle}_{F,j,i}$ is the i-th component of $\boldsymbol{\beta}^{\langle 1\rangle}_{F,j}$; and the superscript "$\langle 1\rangle$" emphasizes that these beta coefficients are obtained in the first step of the SAGLasso procedure.

This first step allows us to screen out a large portion of candidate predictors within each group.[9] In total, only 39 out of 131 series remain and have non-zero coefficients on their contemporaneous and/or lagged values after the adaptive lasso is applied; the number of the selected macro variables is only 58 out of 917 (131×7). Let $\widehat{X}^{\langle 1\rangle}_j, j = i, \ldots, viii$, denote the set of macro variables, in group j, that survive from the first stage.

In the second step, we select those relevant $\widehat{X}^{\langle 1\rangle}_j$ using group lasso. Yield PCs are included as control variables as before. The results from the group lasso show that the coefficients of groups i, iv, v, vi, and viii are shrunk to exactly zero; particularly, group vi (bond and FX) is not selected as a result of controlling for yield factors. For each of the three selected groups—labor market (group ii), housing (group iii), and price indices (group vii)—the group lasso solution obtained from Eq. (19) in Appendix C yields its corresponding group macro factor:

$$\widehat{g}_{\mathrm{j}} = \widehat{X}^{\langle 1\rangle}_{\mathrm{j}} \widehat{\beta}^{\langle 2\rangle}_{\mathrm{j}}, \qquad \mathrm{j} = ii, iii, vii, \tag{4}$$

where j denotes the index of group j whose beta coefficient in step two, $\widehat{\beta}^{\langle 2\rangle}_{\mathrm{j}}$, is not zero. For ease of reference, we relabel $\{\widehat{g}_{\mathrm{j}}\}$ as $\{\widehat{g}_h; h = 1, 2, 3\}$. They each have a clear economic interpretation by

[9] For instance, consider the largest group, the "labor market," that originally contains 32 series and thus 32×7 (=224) variables. Column 7 of Table A.1 indicates that only five series (out of 32), #41, #44, #46, #48, and #49, are selected and that only 11 out of the original 224 variables are selected, including lag-5 and lag-6 of #41; #44 along with its lag-1, lag-2, and lag-3; #46 along with its lag-2; #49 along with its lag-2; and #49 itself.

construction and represent the employment, housing, and inflation factors, respectively.

Unlike inflation and employment, which are commonly incorporated in MTSMs and are well motivated by certain equilibrium term structure models (e.g., Wachter 2006), the housing sector has received little attention in the term structure literature. Given that $\widehat{g}_2$ is a reflection of the share of aggregate consumption devoted to housing, the link between our housing factor and the term premium may be motivated using the idea of Piazzesi, Schneider, and Tuzel (2007) that the expenditure share on housing can drive the equity risk premium.

Note that each of $\{\widehat{g}_\mathfrak{h}\}$ is parsimonious: $\widehat{g}_1$ includes 5 series (11 variables); $\widehat{g}_2$ 8 series (13 variables); and $\widehat{g}_3$ 6 series (6 variables). In total, out of the original 131 series (917 variables), we identify 19 series (30 variables) associated with labor market, housing, and prices that have strongest connection with bond risk premia but the least overlap with yield PCs. Moreover, 21 selected variables (out of 30) are lagged, indicating that many series have a lagged effect on bond risk premia. In particular, certain types of shocks to consumer prices or the labor market seem to require a long lag to manifest their impact on the bond market. The SAGLasso method allows us to select those important lagged variables and capture their lag effect on bond risk premia (e.g., $\widehat{g}_3$ includes no current CPI and PPI variables).

Figure 1 provides a visualization of the selected macro variables. To illustrate the words most relevant to bond return prediction, the word cloud font is drawn proportional to the number of selected macro series (including lagged variables) in which the word appears. The most notable finding is that new housing units started and authorized are highly informative about bond risk premia. In addition to the group level information, the word cloud also reveals the most important subsectors within each selected group. For example, housing market condition in the west and northeast states seem to play a more important role than that in the midwest. Also, commodity price indices appear to be more useful than more general price indices for bond return prediction.

For purposes of forecasting, term structure modeling, and model comparison, we construct a

single aggregate macro predictor using the aforementioned three group factors:

$$\widehat{G} \equiv \sum_{h=1}^{3} \widehat{g}_h. \tag{5}$$

We refer to this predictor as the SAGLasso (single) macro factor hereafter. Note that this factor is a linear combination of only 30 macro variables belonging to merely 19 different series, yet it has strong predictive power for bond risk premia as shown below.

4.3 A Recursively Constructed SAGLasso Factor

The SAGLasso factor constructed in Section 4.2 is based on the full sample and is an unconditional/static factor. Below we construct a dynamic SAGLasso factor recursively. To avoid forward-looking bias, we estimate everything using only the information available at the time of the forecast; namely, we recursively re-estimate *both* factors and parameters when the new information becomes available. We denote a recursively constructed factor by a tilde (e.g., $\widetilde{G}$) to differentiate it from its unconditional counterpart, denoted by a hat (e.g., $\widehat{G}$).

Suppose we want to construct $\widetilde{G}$ at month t based on observations from $t-R$ to $t-1$ and use the predictor to help forecast one-step-ahead annual excess bond returns, where $R > 1$ denotes the number of months included in the training period. Namely, in month $t = R$ we have the following information set of monthly observations available: $\mathcal{F}_R = \{X_t, \{rx_{t,t+12}^{(n)}, 2 \leq n \leq 5\}, t = 1, \ldots, R\}$.

To examine the importance of macro variables over time, we focus on rolling-window estimations.[10] That is, we construct $\widetilde{G}$ at, say, t+1 using observations from t-R+1 to t. We use $R = 240$ (a 20-year training period) in this exercise. Figure 3 depicts the importance of individual macro variables over time. From the rolling-window prediction at time t, we extract coefficients of standardized macro variable k and their lagged values $\beta_{k,l,t}(1 \leq k \leq 131, 0 \leq l \leq 6)$, and map their

[10] See, e.g., Lewellen (2015) who uses a 10-year rolling window to form OLS-based forecasts of individual stock returns and finds that the importance of many characteristics diminishes over time. The procedure using an expanding window to construct $\widetilde{G}$ has higher stability than that using the rolling window: $\widehat{g}_1, \widehat{g}_2$ and $\widehat{g}_3$ are the only groups selected. At the individual level, variables #42 (belonging to "labor market") and #53 (belonging to "housing sector") are the only new variables selected in certain months (and not included in the unconditional $\widehat{G}$ factor). The predictive power of $\widetilde{G}$ with the expanding window is closely comparable to that with the rolling-window.

norm $\sqrt{\sum_l \beta_{k,l,t}^2}$ to the color gradients displayed on the right side of the figure. At the group level, the selection results are fairly stable over time: The labor, housing, and inflation groups are selected in most months. The only exception is the 2002–2005 period, during which macro variables in housing and inflation groups diminished in importance and a couple of variables on industrial production are selected instead.[11] At the individual level, the selected macro series are consistent with the results in Figure 1. Within the labor market group, nonfarm payrolls in the manufacturing and financial sectors play crucial roles in bond return predictions. In the inflation group, the commodity price index appears the most prominent determinant of bond risk premiums.

4.4 Predictive Power of the SAGLasso Factor

4.4.1 In-Sample Evidence

Figure 2 plots the SAGLasso factor (in blue) and excess returns on the five-year bond (in orange) in the full sample period, where shaded areas indicate the periods designated by the National Bureau of Economic Research (NBER) as recession periods. As expected, $\widehat{G}$ captures the countercyclic component in risk premia and leads movements in the realized bond returns. Indeed $\widehat{G}$ generally starts rising at the early stage of economic downturns and peaks during recessions; accordingly, excess bond returns follow and tend to peak toward the end of (or even after) recessions.

Panel A of Table 1 presents results on the in-sample predictive power of $\widehat{G}$, for 2-, 3-, 4-, and 5-year bonds, over the full sample. Test statistics are reported for two different standard errors: Hansen and Hodrick (1980) GMM (in parentheses) and Newey and West (1987) (in brackets).[12] Columns (1)–(4) show that $\widehat{G}$ alone has significant predictive power for excess returns, with the R^2 ranging from 0.35 for the 2-year bond to 0.39 for the 5-year one. Columns (5)–(20) indicate that the

[11] Given that the housing market boom after the early 2000s recession makes the share of housing consumption less of a concern, it is unsurprising that variables in the housing sector become less important in this period. By the same logic, the decline in the importance of inflation indices can be attributable to the stable inflation uncertainty in 2000s (e.g., Wright 2011).

[12] In an earlier version, we also report the t-statistics with Hodrick (1992) 1B covariance estimator, which is constructed using the approximate method of Wei and Wright (2013). The results for $\widehat{G}$ are qualitatively similar, but other return predictors tend to lose their significance with the Hodrick standard errors.

significance of $\widehat{G}$ is robust to each of the following four factors: (a) a modified LN09 factor ($\widehat{LN}^m$), (b) the Cochrane and Piazzesi (2005) forward-rate factor (CP), the Duffee (2011) hidden factor ($\widehat{H}$), and the convergence gap ($\widehat{CG}$) defined by Berardi et al. (2021).[13] The $\widehat{G}$ factor, however, does not completely subsume any of these four factors. The main reason is that whereas $\widehat{G}$ is a pure macro factor by construction, $\widehat{LN}^m$ includes Treasury and FX variables (group vi), $\widehat{CG}$ exploits information in the Federal Funds rate market, and both $\widehat{CP}$ and $\widehat{H}$ are purely yield-curve-based factors. For example, $\widehat{G}$ does not subsume $\widehat{CG}$ for the 2-year bond in the bivariate regression. This result is intuitive given that by construction, $\widehat{CG}$ is expected to be most informative about short-term bond premiums while $\widehat{G}$ is trained on the aggregate bond market returns rather than a specific-maturity bond. As another example, if yield PCs are not controlled for in the second step of the construction of $\widehat{G}$, then the resultant $\widehat{G}$ subsumes $\widehat{LN}^m$ (Huang and Shi 2010).

Panel B reports the results for 2-, 5-, 7-, and 10-year bonds for the post-1984 subsample. While the results on $\widehat{G}$ are generally similar to their counterparts in panel A, the predictive power of the other return predictors all becomes weaker except for $\widehat{CG}$. For instance, $\widehat{G}$ now subsumes $\widehat{LN}^m$ under the HH correction, but $\widehat{CG}$ has increased values of both the t-statistics and incremental R^2s.

In summary, Table 1 shows that $\widehat{G}$ has both significant unconditional and conditional predictive power for bond risk premia. Additionally, $\widehat{G}$ subsumes other macro-based predictors post-1984. In Internet Appendix IA.B, we also conduct in-sample spanning tests and find that both H_0^{S1} and H_0^{S2} are overwhelming rejected.

4.4.2 Out-of-Sample Accuracy

We next examine the out-of-sample performance of the SAGLasso factor, focusing on its incremental power above and beyond yield-curve factors.

[13] In an untabulated analysis, we also consider the output gap factor (*gap*) of Cooper and Priestley (2009); the new-order factor (*NOS*) of Jones and Tuzel (2013); the Cieslak and Povala (2015) "cycle" factor based on yield curves and inflation; and a realized jump-mean factor constructed by Wright and Zhou (2009) (the latter two for the post-1984 sample only). We find that $\widehat{G}$ subsumes *gap* and *NOS* and is not driven out by the other two factors. Chernov and Mueller (2012) uncover a hidden factor that captures inflation expectations as well as bond risk premia; however, this "survey" factor is present only in models estimated with survey-based information.

We divide the sample into training/estimating and out-of-sample (testing) portions. The former consists of $R > 1$ observations. We use fixed rolling-windows with $R = 240$ ($R = 180$) for the full sample (sub-sample) analysis. If P denotes the number of one-step-ahead predictions, then $T = R+P+12$, where T is the total number of observations of macro series. We construct $\widetilde{G}$ recursively month by month using only information available at the time of estimation as described in Section 4.3. Similarly, we recursively re-estimate the yield-curve factors $PC^o_{1-3,t}$ and $PC_{1-5,t}$, whose dynamic versions are denoted $\widetilde{PC}^o_{1-3,t}$ and $\widetilde{PC}_{1-5,t}$.[14]

Given the dynamic macro and yield-curve factors, we form our out-of-sample tests of H_0^{S1} as follows: Consider a "restricted" benchmark model and an "unrestricted" model, where the former is the return forecasting model solely based on $\widetilde{PC}^o_{1-3,t}$ and the latter includes $\widetilde{PC}^o_{1-3,t}$ and $\tilde{G}_t$. Given this pair of nested specifications, we can obtain their time series of realized forecast errors over the entire (out-of-sample) testing period and then conduct a model comparison. In other words, the statistical significance of $\tilde{G}$'s incremental predictive power can be assessed by testing the null hypothesis that the restricted model encompasses the unrestricted one. We form tests of H_0^{S2} similarly by replacing $\widetilde{PC}^o_{1-3,t}$ with $\widetilde{PC}_{1-5,t}$.

Panel A of Table 2 accesses the out-of-sample performance of $\tilde{G}$ with three metrics: the out-of-sample R^2 (Campbell and Thompson 2008) along with its incremental changes due to $\tilde{G}_t$ (R^2_{oos} and ΔR^2_{oos}), and two encompassing tests for nested models—the Ericsson (1992) ENC-REG and Clark and McCracken (2001) ENC-NEW tests.[15] The R^2_{oos} levels of $\tilde{G}_t$ show that $\tilde{G}_t$ alone captures nontrivial real-time information on bond risk premiums. Also, the R^2_{oos} increases with the bond maturity. In fact, the R^2_{oos} for the 2-year bond is substantially lower than that for the 5-year

[14] To reduce the computational burden, we estimate the parameters in model $YTSM(5)$ only once using the full sample and then extract $\widetilde{PC}_{1-5,t}$ using filtering (not smoothing) from the estimated model. That is, $\widetilde{PC}_{1-5,t} = \widehat{PC}_{1-5,t}$ in Section 4.4.2. Using $\widetilde{PC}_{1-5,t}$, however, is biased against the predictive power of $\tilde{G}_t$. Indeed, we find that using $(\widetilde{PC}^o_{1-3,t}, \widehat{PC}_{4-5,t})$ instead of $\widehat{PC}_{1-5,t}$ slightly strengthens $\tilde{G}_t$'s predictive power (untabulated).

[15] The precise asymptotic distribution of the test statistics in these two tests depends on the asymptotic ratio of P/R, denoted by $\pi \equiv \lim_{P,R\to\inf} P/R$. The Ericsson test critical values from a standard normal distribution are conservative if $\pi > 0$. Given that $\pi \geq 1$, the simulation results of Clark and McCracken (2001) show that the 95% critical value ranges from 1.584 to 2.685 for testing one additional predictor.

(10-year) bond in the full sample (subsample).[16].

Panel A1 (A2) shows that incorporating $\tilde{G}_t$ into the restricted model based on $\widetilde{PC}^o_{1-3,t}$ ($\widetilde{PC}_{1-5,t}$) and a constant improves the model performance substantially in either the full or sub sample. First, both the ENC-REG and ENC-NEW test statistics greatly exceed their asymptotic critical values, regardless how the asymptotic ratio of P/R is specified, thereby rejecting both H_0^{S1} and H_0^{S2}. Second, including $\tilde{G}_t$ also raises R^2_{oos} substantially. For instance, when $\widetilde{PC}^o_{1-3,t}$ is augmented with $\tilde{G}_t$, ΔR^2_{oos} ranges from 0.271 for the 5-year bond to 0.349 for the 2-year bond in the full-sample analysis. Note that the high values of ΔR^2_{oos} here are partially attributable to the negative R^2_{oos} values under the restricted models. To summarize, panel A shows that the improvement in forecasting accuracy due to $\tilde{G}$ is statistically significant.

4.4.3 Economic Values

We now examine economic gains of $\widetilde{G}$'s out-of-sample predictive power. We follow Campbell and Thompson (2008) and assess a mean-variance investor's utility gains from trading on $\widetilde{G}$ against a benchmark. The investor is assumed to dynamically allocate her portfolio between an N-year bond ($N \geq 2$) and a one-year bond (the risk-free asset) at a monthly basis, based on the standard optimal (timing) strategy (e.g., Thornton and Valente 2012). Given her risk aversion coefficient (γ) and the N-year bond return volatility at time t, the investor implements the strategy based on her out-of-sample forecasts of the N-year bond risk premium.

We consider three return predictors: $\widetilde{G_t}$, $\widetilde{PC}^o_{1-3}$, and $\widetilde{PC}^o_{1-3,t}$+$\tilde{G}$. The timing strategies based on these predictors are denoted $\mathcal{S}^G$, $\mathcal{S}^Y$, and $\mathcal{S}^{G+Y}$, respectively. In addition, we consider a buy-and-hold strategy, denoted $\mathcal{S}^{BH}$. We then compare $\mathcal{S}^G$ against $\mathcal{S}^{BH}$, as well as $\mathcal{S}^{G+Y}$ against $\mathcal{S}^Y$, to examine incremental welfare gains due to $\widetilde{G}$. Specifically, we calculate the certainty equivalent return (CER) values for each month in the testing sample and then estimate the following regression: $u_{g,t}-u_{0,t} = \nu+\varepsilon_t$, where $u_{g,t}$ and $u_{0,t}$ represent realized utilities generated by strategies $\mathcal{S}^G$ and $\mathcal{S}^{BH}$

[16]Bianchi et al. (2021) find that the performance of their macro factors is also relatively weak for short-term bonds.

($\mathcal{S}^{G+Y}$ against $\mathcal{S}^{Y}$), respectively. To examine whether the incremental utility gains are significant or not, we test the null hypothesis that $\nu = 0$ (denoted H_0^{ν}) using a variant of the Diebold and Mariano (1995) test, proposed by Harvey et al. (1997), that accounts for autocorrelation in the forecasting errors.

Panel B of Table 2 reports the annualized CER values along with the corresponding p-values for H_0^{ν} (in angel brackets) with $N = 2, 3, 4, 5$ for the full sample or $N = 2, 5, 7, 10$ for the post 1984 subsample. In each panel we consider two risk version levels: $\gamma = 3$ as adopted by Campbell and Thompson (2008) and Gu et al. (2020), and $\gamma = 5$ as adopted by Thornton and Valente (2012) and Bianchi et al. (2021). We also follow these studies to limit the portfolio weight on the N-year bond to lie between 0 and 150%.

Results for $\mathcal{S}^{G}$ vs. $\mathcal{S}^{BH}$, reported in Panel B1, indicate that the out-of-sample predictive power of $\tilde{G}$ can generate sizable welfare benefits relevant for investors. For example, in the case of $\gamma = 5$ with $N = 5$, $\mathcal{S}^{G}$ leads to certainty equivalent gains of 8.62% (4.05%) relative to $\mathcal{S}^{BH}$ for the full (post-1984) sample. Campbell and Thompson (2008) show that the investor's welfare gain depends on the relative magnitude of predictive R^2 and the buy-and-hold Sharpe ratio. Since the R^2_{oos} values of $\tilde{G}$ increases with the bond maturity and the Sharpe ratio decreases with the maturity, it is not surprising to find that CER values become greater as the bond maturity increases.

Results for $\mathcal{S}^{G+Y}$ vs. $\mathcal{S}^{Y}$, reported in panel B2, show that the hypothesis H_0^{ν} is rejected at the 5% significance level in all but one case (with $N = 2$ and $\gamma = 5$). In other words, incorporating $\tilde{G}$ into the out-of-sample forecasting of the bond risk premium can lead to significant utility gains relative to trading on $\widetilde{PC}^{o}_{1-3}$ alone. Since these utility differences have the units of expected annualized return, they can be roughly interpreted as the differences in portfolio management fees. We find that a mean-variance investor with $\gamma = 3$ is prepared to pay extra 43-113 bps per year to exploit the additional information as contained in factor $\tilde{G}$.

To summarize, Section 4 provides strong evidence against H_0^{S1} and H_0^{S2}. It also shows that

rejection of the these two hypothesis carries significant economic values.

4.4.4 Additional Evidence

We further examine the predictive power of the SAGLasso factor in Internet Appendix IA.B and summarize the main findings here.

Given that $\widehat{LN}^m$ is constructed using the same set of 131 macro series and includes all 131 series as well as squares and cubes of these macro variables, $\widehat{LN}^m$ serves as a natural benchmark for $\widehat{G}$ (a linear combination of 19 series and some of their lagged variables). We find that $\widehat{G}$ shows stronger predictive ability than $\widehat{LN}^m$ in both in-sample and out-of-sample analyses.

As mentioned before, the set of 131 macro series we use is adjusted for both data revisions and publication lags. One relevant question is the impact of these two adjustments on bond return predictability. We find that the return predictability evidence based on $\widehat{G}$ is not sensitive to the vintage of macro data used. In contrast, publication lags pose much greater "danger" than data revisions in forecasting future bond returns based on macro variables, at least in our sample. This problem can be mitigated straightforwardly, however, since it is easier to make an adjustment for publication lags than to figure out preliminary macro data releases and adjust for data revisions.

To better understand the source of the predictive power of the SAGLasso factor ($\widehat{G}_t$), we also examine properties of its three components: the employment ($\widehat{g}_{1t}$), housing ($\widehat{g}_{2t}$), and inflation ($\widehat{g}_{3t}$) factors. As expected, $\widehat{g}_{1t}$, $\widehat{g}_{2t}$, and $\widehat{g}_{3t}$ all have low correlations with the yield curve factors; as a result, $\widehat{G}$ is weakly correlated with $PC^o_{1-3,t}$ and hardly correlated with $PC_{4,t}$ and $PC_{5,t}$. The three group factors also show significant predictive power, both individually and jointly. Following JPS, we also examine the relative importance of the three group factors across bond maturity. Our results indicate that relatively speaking, among the three group factors, $\widehat{g}_{1t}$ is the most important, followed by $\widehat{g}_{3t}$, and then by $\widehat{g}_{2t}$, regardless of the bond maturity.[17]

[17] Bianchi et al. (2021) consider more categories and find that variables related to the stock and labor market (the output & income and orders & inventories) are more important for the short-end (long-end) of the yield curve. Note that the aggregate bond market is used to train $\{\widehat{g}_h\}$.

The SAGLasso algorithm is implemented using 131 macro variables along with six of their lags. One question that arises is: Are lags of macro variables are essential to the predictive power of the SAGLasso factor? If yes, what is the optimal number of lags to be included in our supervised learning? We repeat the baseline analysis using the 131 macro variables along with N_L of their lags, where $N_L = 0, 3, 9, 12$. We find that the evidence of the return predictability is robust to the use of no lags ($N_L = 0$). Nonetheless, our results suggest that the SAGLasso factor constructed using the 131 macro variables along with 3 or 6 of their lags has the best performance in both the in-sample and out-of-sample predictions. This finding reflects a trade-off between including more information in the supervised learning and imposing a denser data structure to enhance the estimation stability. While the baseline SAGLasso factor (with $N_L = 6$) seems to capture more information on long-term bond premiums, the alternative SAGLasso factor with $N_L = 3$ outperforms for short-term bonds.

5 The SAGLasso factor and the Spanning Controversy

As an important application of the SAGLasso factor, we revisit the spanning controversy in this section. We focus on the three main aspects of the controversy. First, whether a macro factor's predictive power is robust to finite samples (see Section 2). Second, whether a macro factor is an unspanned pricing factor in an MTSM. Third, whether or not a macro factor's temporal variation can be captured by the yield curve. We show that the SAGLasso factor can address all three aspects of the controversy by using the dynamic term-structure modeling framework.

5.1 The Modeling Framework

Following JPS, we assume that all risks in the economy are encompassed by an $\mathcal{N}$-dimensional state vector $X_t = (\mathcal{P}_t, F_t)$, where $\mathcal{P}_t$ denotes $\mathcal{L}$ linear combinations of (noise-free) zero yields and the ($\mathcal{N}$-$\mathcal{L}$)-vector F_t represents macro factors as before. The short rate is an affine function of X_t:

$$r_t = \delta_0 + \delta_1' X_t \;=\; \delta_0 + \delta_{1p}' \mathcal{P}_t + \delta_{1f}' F_t. \tag{6}$$

The dynamics of X_t under the risk-neutral measure $\mathbb{Q}$ are assumed to follow a Gaussian process:

$$X_t = \begin{bmatrix} \mathcal{P}_t \\ F_t \end{bmatrix} = \begin{bmatrix} \mu_p^{\mathbb{Q}} \\ \mu_f^{\mathbb{Q}} \end{bmatrix} + \begin{bmatrix} \Phi_{pp}^{\mathbb{Q}} & \Phi_{pf}^{\mathbb{Q}} \\ \Phi_{fp}^{\mathbb{Q}} & \Phi_{ff}^{\mathbb{Q}} \end{bmatrix} \begin{bmatrix} \mathcal{P}_{t-1} \\ F_{t-1} \end{bmatrix} + \Sigma_x \epsilon_{x,t}^{\mathbb{Q}}, \qquad \epsilon_t^{\mathbb{Q}} \sim MVN(0, I). \tag{7}$$

It follows from Duffie and Kan (1996) that the yield of an m-period zero-coupon bond is

$$y_t^{(m)} = A_m + B_m' X_t, \tag{8}$$

where the expressions for A_m and B_m are given in Internet Appendix IA.C.1. The market price of risk follows the "essentially affine" structure of Duffee (2002):

$$\Sigma\Lambda_t = \mu_x^{\mathbb{P}} - \mu_x^{\mathbb{Q}} + (\Phi^{\mathbb{P}} - \Phi^{\mathbb{Q}})X_t \; = \; \lambda_0 + \lambda_1 X_t, \tag{9}$$

where $\{\mu^{\mathbb{P}}, \Phi^{\mathbb{P}}\}$ are the $\mathbb{P}$-measure counterparts of $\{\mu^{\mathbb{Q}}, \Phi^{\mathbb{Q}}\}$.

5.2 Finite Sample Analysis

The statistical inference done in Section 4.4 is based on asymptotic distributions. We now examine H_0^{S1} and H_0^{S2} using a finite-sample analysis. This is necessary because, first, our dependent variables involve overlapping observations by construction, and secondly, the first and second PCs of yield curves are highly persistent in our sample, with first-order autoregressive coefficients (ACF) of 0.99 and 0.94, respectively (while the ACF of the SAGLasso factor is only 0.82). Below we first specify the underlying data-generating processes (DGPs) for H_0^{S1} and H_0^{S2} within the framework described in Section 5.1. We then construct finite-sample distributions of test statistics from return-forecasting regressions and conduct finite-sample inference based on such distributions.

5.2.1 Data-Generating Processes for Null Hypotheses

DGPs under H_0^{S1} or H_0^{S2} impose no restrictions on model parameters and allow them to be estimated freely. That is, as long as the $\mathcal{N} \times \mathcal{N}$ yield loading matrix $\mathcal{B} \equiv (B_{m_1}, \ldots, B_{m_\mathcal{N}})'$ is invertible, the fraction of variations in term premia that are associated with macro factors is also attributable

to certain linear combinations of these yields. This type of MTSMs are referred to as spanned models and denoted by $SM(\mathcal{L},\mathcal{N})$. If $\mathcal{B}$ is not invertible, then the model is no longer spanned.

Given that $F_t = G_t$, the DGP for H_0^{S1} is model $SM(2,3)$. To see why, suppose that yield PCs are defined in terms of k zero-coupon bonds with maturities $\mathcal{M} = \{m_1, \ldots, m_k\}$ as follows:

$$PC_{1-\mathcal{N},t} = WY_t^{\mathcal{M}} \equiv W(\mathcal{A}_{\mathcal{M}} + \mathcal{B}'_{\mathcal{M}} X_t), \quad W \in \mathbb{R}^{\mathcal{N}\times\mathcal{N}}.$$

Since $SM(2,3)$ is a spanned model, rank $(\mathcal{B}_{\mathcal{M}}) = \mathcal{N} = 3$. The resultant invertibility of $W\mathcal{B}'_{\mathcal{M}}$ implies

$$E_t\left(rx_{t,t+12}^{(12n)}\right) = \text{constant} + \psi'_{12n,12}(W\mathcal{B}'_{\mathcal{M}})^{-1} PC_{1-3,t}, \tag{10}$$

where $\psi_{m,12} = mB'_m - (m-12)B'_{m-12}(\Phi^{\mathbb{P}})^{12} - 12B_{12}$ for $m > 12$. This result means that G_t has no incremental predictive power for annual excess returns in the presence of $PC_{1-3,t}$, consistent with H_0^{S1}. Similarly, the DGP for H_0^{S2} is model $SM(4,5)$.

At the heart of Eq. (10) is the theoretical spanning of G_t by any three zero yields. In other words, as long as $k \geq \mathcal{N}$, the covariance matrix of $Y_t^{\mathcal{M}}$ (stacked bond yields) has a rank of 3. However, empirically the sample covariance matrices are nonsingular regardless of the choice of maturities $\mathcal{M}$. The standard interpretation in the literature is that observed yields (denoted $Y_t^{o\mathcal{M}}$) are contaminated by small transitory noise, modeled as idiosyncratic "measurement error" (representing a catch all term for model misspecification and other imperfections) as follows:

$$Y_t^{o\mathcal{M}} = \mathcal{A}_{\mathcal{M}} + \mathcal{B}'_{\mathcal{M}} X_t + \eta_{yt}, \quad \eta_{yt} \sim MVN(0, \sigma_{\eta_y}^2 I). \tag{11}$$

The presence of η_{yt} is also important in terms of accommodating hidden yield factors in spanned models with $\mathcal{N} > 3$. For instance, consider model $SM(4,5)$, where $PC_{1-5,t}$ fully determine the term premia and absorb the role of G_t. If at least five zero yields (or their linear combinations) are assumed to be measured without error, the full-rank $\mathcal{B}'_{\mathcal{M}}$ indicates that the entire state vector can be perfectly extracted from the five yields. Consequently, H_0^{S2} degenerates into a version of H_0^{S1} that involves more than three yield PCs. Alternatively, if measurement error is ubiquitous, it becomes difficult to extract higher-order PCs, say, $PC_{4,t}$, from the cross section of yields. As

such, Eq. (11) opens up the possibility that bond risk premia contain a component attributable to higher-order PCs, yet hidden from the observed yield curve—namely, a hidden factor.

5.2.2 Finite-Sample Inference

This subsection reports finite-sample properties of test statistics under H_0^{S1} or H_0^{S2}, whose underlying DGPs are $SM(2,3)$ and $SM(4,5)$, respectively. We estimate these spanned models using the full-sample zero-coupon yields with maturities $\mathcal{M} = \{0.25, 1, 2, 3, 4, 5\}$ to generate samples over the period 1964–2014, or using extended Fama-Bliss zero yield data with $\mathcal{M} = \{0.5, 1, 2, 3, 4, 5, 7, 10\}$ to generate samples for the post-1984 period.

As the inference about H_0^{S2} requires all yields to be measured with errors, we implement the model estimation with maximum likelihood using the Kalman filter. To facilitate the interpretation of the sources of risk compensation, we normalize yield-based state variables $\mathcal{P}_t$ to the first $\mathcal{L}$ PCs of zero yields; namely, $X_t = (PC_{1-\mathcal{L},t}, G_t)$. This rotation also offers OLS-based starting values in the estimation of $\mathbb{P}$-dynamics of X_t. When estimating $\mathbb{Q}$-measure parameters, we rotate X_t to X_t^*, a state vector that satisfies the canonical form of Joslin, Le, and Singleton (2013).[18]

Under each spanning hypothesis, we generate 5,000 artificial data sets from its underlying DGP estimated with the full or post-1984 sample. In the in-sample analysis, we obtain the distributions for two t-statistics (based on HH and NW standard errors, respectively) and R^2.[19] In the out-of-sample analysis we consider the ENC-REG and ENC-NEW tests and R^2_{oos}.[20] We calculate the 5% critical value and p-value for each set of statistics, the latter being defined as the frequency of bootstrap replications in which the test statistics are at least as large as in the real data.

[18] In other words, instead of directly estimating parameters in Eqs. (6) and (7), we estimate another (and shorter) parameter vector $\Theta_M^{\mathbb{Q}}$ (defined in Internet Appendix IA.C.1) that encompasses all bond pricing information.

[19] We do not consider the t-statistic based on the Hodrick (1992) standard errors here because it tends to under-reject the null. Also, Ang and Bekaert (2007) show that it has desirable small-sample properties.

[20] In our baseline finite-sample inference, there is no distinction between the in-sample factor $\widehat{G}_t$ and the real-time factor $\widetilde{G}_t$. To make the out-of-sample inference truly out of sample, we perform full-scale simulations in which the time series of 131 individual macro variables are generated together with the $\mathcal{N}$-$\mathcal{L}$ yield factors. In each trial, the SAGLasso estimator is implemented on the generated macro variables to construct macro factors $\widehat{G}_t$ and $\widetilde{G}_t$. These re-simulated $\widehat{G}_t$s and $\widetilde{G}_t$s are then used to infer the finite-sample distribution of test statistics. This exercise guards against the data mining concerns being translated into the finite-sample analysis. Unreported results indicate that the properties of test statistics under the full-scale simulations are similar to those under our baseline simulations.

Panel A of Table 3 reports finite-sample properties of test statistics for the full sample. Note from panels A1 (in-sample) and A2 (out-of-sample) that small-sample distortions appear more severe under H_0^{S1}. For in-sample t-statistics, the "true" 5% critical value ranges from 3.46 to 4.47, depending on the bond maturity and standard errors used; for ΔR^2 (the incremental in-sample R^2 due to G_t), the 95th percentile of its small-sample distribution is higher than 9%. However, all of these critical values are substantially lower than actual statistics obtained from our data sample. Similarly, note from panel A2 that there is strong evidence against H_0^{S1}. In particular, all statistics have bootstrapped p-values less than 1%. Also, the critical value of ΔR^2_{oos} ranges from 11.7% for the 5-year bond to 13.6% for the 2-year bond. Results reported in panels A3 (in-sample) and A4 (out-of-sample) of Table 3 illustrate that under H_0^{S2}, small-sample distributions of test statistics show even greater deviations from their asymptotic distributions. For instance, the critical value for the HH t-statistics under H_0^{S2} (panel A3) is at least 0.8 higher than its counterpart under H_0^{S1} (panel A1), with the biggest difference of 1.29 ($= 4.75 - 3.46$) for the 5-year bond. For out-of-sample tests, the ENC-REG critical value is 4.02∼4.36, and the ENC-NEW critical value can be as high as 52.18 in small samples (panel A4), but the critical values are still not large enough to overturn the asymptotic analysis-based rejection of H_0^{S2} concluded in Section 4.4.2.

We find similar results for the post-1984 sample (panel B of Table 3), although statistics estimated from the subsample are subject to less severe distortions than those from the full sample. Particularly, the asymptotic analysis-based evidence against H_0^{S1} and H_0^{S2} post 1984 (panel B of Table 2) is robust to small samples.

Overall, we draw three conclusions from Table 3. First, small-sample bias tends to decrease with the bond maturity. Second, the asymptotic analysis-based evidence against H_0^{S1} and H_0^{S2} (Table 2 and also Internet Appendix IA.B) is too strong to be overturned. Third, results on descriptive statistics show that none of the 5,000 artificial samples are able to generate a ΔR^2 or ΔR^2_{oos} that exceeds the actual incremental R^2.

We present more robustness analyses in the Internet Appendix. Section IA.D shows that model $SM(2,3)$ provides a more robust test of H_0^{S1} than does the DGP proposed in BH. Section IA.E conducts the Ibragimov and Müller (2010) test of H_0^{S1} and H_0^{S2} that is robust to heteroscedasticity, autocorrelation, and structural breaks, and finds that among the five yield factors and the SAGLasso factor, the latter is the only robust bond return predictor. Finally, Section IA.F examines an alternative version of H_0^{S2} where the conditioning variable Z_t is the "cycle" factor of Cieslak and Povala (2015) given that this factor is spanned. We find that this hypothesis is rejected as well.

To summarize, the results from our finite-sample analysis strongly reject the two spanning hypotheses, suggesting that it is very unlikely for a spanned MTSM to account for the additional predictive power of the SAGLasso factor as observed in our sample.

5.3 Testing the Macro-Unspanning Hypothesis

The rejection of the spanning hypotheses with $F_t = \widehat{G}_t$ implies that MTSMs incorporating $\widehat{G}_t$ may be preferable to "yields-only" term structure models (YTSMs), say, for term premium inference. Then a follow-up question is: Should $\widehat{G}_t$ be used as a bond-pricing factor in an MTSM and if yes, is $\widehat{G}_t$ a spanned pricing factor? We address this question by formulating and testing the "macro-unspanning hypothesis" (MUH), which intuitively says that in spite of its predictive power for bond risk premia, $\widehat{G}_t$ is not a spanned pricing factor.

5.3.1 The Macro-Unspanning Hypothesis

In the MTSM framework described in Section 5.1, the MUH (arising from the conditions specified in JPS and BR for unspanned macro risks) can be stated as follows:

$$H_0^{US}: \quad \delta_{1f} = 0 \quad \text{and} \quad \Phi_{pf}^{\mathbb{Q}} = 0. \tag{12}$$

Under these restrictions, the short rate depends only on $\mathcal{P}_t$ ($\mathcal{L}$ linear combinations of zero yields), and the $\mathbb{Q}$-dynamics of F_t as represented by $\{\mu_f^{\mathbb{Q}}, \Phi_{fp}^{\mathbb{Q}}, \Phi_{ff}^{\mathbb{Q}}\}$ are not identifiable without information from other asset markets. It follows that only risks of yield PCs are priced in the Treasury market.

Namely, the one-period risk premium, $\Sigma\Lambda_t$, given below, is $\mathcal{L}$-dimensional:

$$\Sigma\Lambda_t = \mu_p^{\mathbb{P}} - \mu_p^{\mathbb{Q}} + \left[\Phi_{pp}^{\mathbb{P}} - \Phi_{pp}^{\mathbb{Q}}\,,\, \Phi_{pf}^{\mathbb{P}}\right] X_t \;=\; \lambda_0 + \lambda_1 X_t. \tag{13}$$

For convenience, such an $\mathcal{N}$-factor MTSM that satisfies H_0^{US} is termed an unspanned model and denoted $USM(\mathcal{L},\mathcal{N})$.

Note that when $\mathcal{L}=3$, H_0^{US} represents the standard version of the MUH: Macro-based forecasts are not spanned by the contemporaneous yield curve (equivalent to the case focused on in BR's likelihood-ratio tests). When $\mathcal{L}>3$, H_0^{US} denotes a more general version that the predictive ability of macro factors is not spanned by the filtration generated by the yield dynamics. We examine both versions and thereby estimate both models $SM(\mathcal{L},\mathcal{N})$ and $USM(\mathcal{L},\mathcal{N})$ with $\mathcal{L}=3,4,5$ in this analysis. To match the data sample used in JPS and BR, we estimate each of these six models using zero yields with $\mathcal{M}=\{0.5,1,2,3,4,5,7,10\}$ over the period 1985–2007.

Note also that H_0^{US} is *not* simply the opposite of H_0^{S1} or H_0^{S2}. First, while H_0^{US} concerns whether a given macro factor with some explanatory power for term premia is a pricing factor, H_0^{S1} and H_0^{S2} focus on whether variables outside of the bond market provide additional explanatory power for bond risk premia. Second, term structure modeling implications from the outcome of testing H_0^{S1} or H_0^{S2} are different from those of testing H_0^{US}. For instance, suppose $\mathcal{N}=5$. Rejecting H_0^{US} implies a rejection of model $USM(4,5)$, where the alternative model is $SM(4,5)$; namely, it is $SM(4,5)$ versus $USM(4,5)$. In contrast, rejecting H_0^{S2} implies that $USM(5,6)$ ought to be used to infer the risk premium component in long-term yields, and accepting H_0^{S2} means that $SM(4,5)$ (or $YTSM(5)$) should be used; that is, it is $SM(4,5)$ versus $USM(5,6)$.[21]

[21] As a result, a test of H_0^{US} corresponds to a test of equal forecast accuracy for non-nested models in the regression setting in Eq. (1). Suppose that $Z_t = PC_{1-5,t}$ and $F_t = G_t$. The question of interest is whether the additional predictive power of G_t is captured by the six yield factors (i.e., $PC_{1-6,t}$) or any other six linear combinations of "true" yields, similar to an encompassing test for comparing non-nested models: $(PC_{1-5,t}, G_t)$ versus $PC_{1-6,t}$.

5.3.2 Statistical Tests of the Macro-Unspanning Hypothesis

We conduct two tests of H_0^{US}. One is a model-based likelihood ratio (LR) test. As there is no analytic expression available for the limiting distribution under H_0^{US}, we compute the critical values of the test statistic based on the approximation method used by BR. However, the approximation is done conservatively, and as a result, this LR test tends to under-reject H_0^{US}.[22] To circumvent this problem and make a more robust inference, we perform another test of H_0^{US} (a model-free test in the spirit of BR) by directly testing the yield loadings on the SAGLasso factor without imposing no-arbitrage restrictions. Given the assumption that all yields are observed with measurement error, we can focus on the loading matrix $\mathcal{B}'_{\mathcal{M}} = (\mathcal{B}'_{\mathcal{L},p}, \mathcal{B}'_{\mathcal{L},f})$ in Eq. (11) in this model-free test. To implement the test, we first estimate Eq. (11) with the OLS and then conduct LR tests of $\mathcal{B}_{\mathcal{L},f} = 0$.

Panel A of Table 4 reports the results from both the model-based (column 2) and model-free (column 3) tests of H_0^{US}, for $\mathcal{L} = \mathcal{N} - 1 = 3, 4, 5$. Note from column 2 that the LR statistics are always smaller than the 10% critical values, $\forall \mathcal{L}$. An unreported decomposition of the log-likelihood function reveals that the difference between $SM(\mathcal{L}, \mathcal{N})$ and $USM(\mathcal{L}, \mathcal{N})$ mainly derives from the $\mathbb{Q}$-likelihood. This result, as documented by BR for $\mathcal{L} = 3$ with two macro factors, is not surprising as the restrictions in H_0^{US} are not placed on the $\mathbb{P}$-dynamics of $USM(\mathcal{L}, \mathcal{N})$. However, our test results show that the improved yield curve fitting of $SM(\mathcal{L}, \mathcal{N})$ over $USM(\mathcal{L}, \mathcal{N})$ is statistically insignificant, in contrast to BR's finding. The p-values reported in column 3 indicate that H_0^{US} is not rejected by the model-free test either at the conventional significance level of 5%, $\forall \mathcal{L}$.

Results in panel A also suggest that the negative effect of excluding $\widehat{G}$ from fitting the yield curve becomes weaker when $\mathcal{N}$ increases. This finding is not surprising: Although the higher-order PCs are considered to be unimportant in explaining cross-sectional variations in yields, they help fit the term structure more or less. Thus, when an additional yield factor is included in the model,

[22] As discussed in BR, while H_0^{US} imposes four zero restrictions for the case of $\mathcal{L} = 3$, a comparison of test statistics with the critical values of a $\chi^2(4)$-distribution would be misleading. Under the approximation adopted by BR (detailed in their Section 3.1), test statistics are evaluated against a χ^2-distribution with $(k - \mathcal{N})(\mathcal{N} + 1) - 1$ degrees of freedom when only one macro variable is used, where k is the number of bonds involved.

the already limited role of G_t in the cross section becomes more redundant.

To summarize, when the SAGLasso factor is used as the sole macro factor of an unspanned model, both the model-based and model-free tests fail to reject the MUH. As mentioned before, the main reason for this finding is that in spite of its strong predictive power for excess bond returns, the SAGLasso variable is weakly correlated with yield PCs and is unspanned (see Section 5.4). See Internet Appendix IA.G for more applications of unspanned models.

5.4 Is the SAGLasso Factor Unspanned?

To examine whether the yield curve can explain the temporal variation in the SAGLasso factor, we follow JPS and regress G_t on $\mathcal{N}$ observed yield PCs:

$$G_t = \gamma_0 + \gamma_1 PC^o_{1-\mathcal{N},t} + \varepsilon_t. \tag{14}$$

To see whether the regression R^2 is low enough to invalidate spanned models, we follow BR and evaluate it against its distribution implied from an $\mathcal{N}$-factor spanned model rather than against unity. To this end, we consider distributions implied by "unconstrained" models as well as "constrained" ones, and also allow for macro measurement error, denoted by η_f with a standard deviation of σ_{η_f}. In contrast, BR focus on unconstrained models with zero η_f. Unconstrained models here refer to MTSMs imposing no constraints on the Sharpe ratio (SR) of bond returns. Such models may imply unrealistic SRs, as noted in Duffee (2010) and Joslin, Singleton, and Zhu (2011). MTSMs with the selected zero restrictions on $\{\lambda_0, \lambda_1\}$ are referred to as constrained models and denoted $CSM(\mathcal{L},\mathcal{N})$ for spanned models and $CUSM(\mathcal{L},\mathcal{N})$ for unspanned models, with $\mathcal{L}$ being the number of yield factors included in the model (see Internet Appendices IA.C and IA.G).

Panel B of Table 4 reports the empirical R^2 value and its 95% confidence interval (in brackets underneath) in column 5, where the interval is based on 5,000 data sets simulated from constrained model $CSM(\mathcal{N}\text{-}1,\mathcal{N})$, estimated with and without macro measurement errors, for $\mathcal{N} = 4, 5, 6$. First, consider the case without macro measurement errors ($\eta_f = 0$), a commonly made assumption in

the macro finance literature (see, e.g., JPS and BR). The results show that $\forall \mathcal{N}$, the empirical R^2 is around 14.5% and outside of its 95% confidence interval with a p-value (defined as the fraction of the simulated samples in which the R^2 is below the value in the actual data) lower than 2.5%. That is, the SAGLasso factor indeed has R^2 values too low to be reconcilable with spanned models. We also evaluate empirical R^2s against their distributions implied from unconstrained models $SM(\cdot)$ and find that the results are similar to those reported in panel B. Since we assume in our model estimation that bond yields are all measured with error, the aforementioned results provide evidence that yield measurement error does *not* account for the large proportion of unspanned macro variation as observed in the real data in our sample.[23]

Next, we assume that $\eta_f \neq 0$. Intuitively, allowing for macro measurement errors would create a further unspanned variation of G_t and thus make it more likely for spanned models to reproduce documented regression evidence. We re-estimate model $CSM(\mathcal{N}\text{-}1, \mathcal{N})$ assuming $\eta_f \neq 0$ and find that the resulting implied R^2 distributions are barely distinguishable from their counterparts with zero η_f. For example, the 95% confidence intervals implied from model $CSM(3,4)$ with and without macro measurement error are $[0.587, 0.769]$ and $[0.593, 0.847]$, respectively (column 5 of Table 4). As a result, even if including η_f shifts the model-implied R^2 distribution to the left, the net impact is minimal; that is, unspanned macro variation observed in our sample cannot be attributed to macro measurement errors either. Behind this finding is the tiny standard deviation of the measurement error in $\widehat{G}_t$: $\widehat{\sigma}_{\eta_f} < 3$ bps for $3 \leq \mathcal{N} \leq 6$. Note that as $\widehat{G}_t$ is standardized under the SAGLasso procedures (Section 4.2), $\widehat{\sigma}_{\eta_f}$ is negligible compared to the total standard deviation of $\widehat{G}_t$.

Panel B of Table 4 also includes the results from a spanning test applicable to macro factors allowed to contain "noise" (Duffee 2013a): if yields span the true state vector, the regression

[23] BR consider regressions similar to Eq. (14) albeit with *GRO* or *INF* as the dependent variable; their simulation results, based on unconstrained models, indicate that adding small yield measurement error makes spanned models capable of generating the appearance of unspanned macro information in the real data. In an untabulated analysis we show that the main reason for such simulation results is, however, that when a macro variable with a low correlation to the yield curve is used as a spanned factor, most variation in this macro factor is captured by high-order yield factors *by construction*; as a result, a spanned model with small yield measurement error can reproduce a large amount of unspanned macro variation even if the macro variable under consideration is unspanned.

in Eq. (14) should produce serially uncorrelated residuals even though the estimated R^2 could substantially deviate from one. The estimated first-order correlation of residuals of the regression is around 0.67, $4 \leq \mathcal{N} \leq 6$ (column 6). Given that the serial correlation of G_t is 0.71, the above result suggests that whatever the regression is missing cannot be explained by white-noise shocks.

Overall, the results of Section 5.4 provide strong evidence that much of the variation in G_t is not captured by the yield curve. This unspanned nature of the SAGLasso factor reinforces our earlier conclusion that it carries term premium information independent of the yield curve. Moreover, this macro variable has very small measurement error even when it is included as a spanned factor in a low-dimensional MTSM.

6 Conclusion

There is no consensus in the literature on whether or not macro variables have incremental predictive power for future excess bond returns over contemporaneous bond yields. However, macro variables considered in the empirical literature are typically standard ones, such as measures of real growth and inflation. These variables either show little unconditional predictive power for bond risk premia or are highly correlated with contemporaneous yields and thus have insignificant conditional predictive power. In this study we construct a new macro variable using Supervised Adaptive Group LASSO (SAGLasso), a machine learning algorithm, from a panel of 917 macro variables (131 macro series along with six of their lags) that are adjusted for both data revisions and publication lags. We show that this new macro variable, termed the SAGLasso (macro) factor, has strong out-of-sample predictive power for bond risk premia conditional on the yield curve. Additionally, this predictability can provide investors with significant economic gains.

Importantly, the SAGLasso factor is parsimonious, intuitive, and easy to interpret. Specifically, it is a *linear* combination of merely 30 selected variables out of 917, and consists of a novel housing factor, an employment factor, and an inflation factor. In addition, in spite of its strong predictive

power, the SAGLasso factor has low correlations with contemporaneous yields by construction; thus, it is a "pure" macro-based bond return predictor.

The SAGLasso macro factor also provides a potential resolution to the spanning controversy in the macro-finance literature. First, the SAGLasso factor is not spanned by contemporaneous yields. Second, in an MTSM with the SAGLasso factor as its sole macro factor, the hypothesis that it is unspanned by the yield factors is not rejected. Third, incorporating the unspanned SAGLasso factor into an MTSM with realistic Sharpe ratios has nontrivial economic benefits. Fourth, the importance of the SAGLasso factor cannot be attributed to measurement errors in yields or itself. Furthermore, its measurement error is small.

To summarize, using a machine learning algorithm we are able to construct a new, parsimonious, and easy-to-interpret macro variable with strong and robust predictive power for bond risk premia. In addition, this new macro factor can potentially help resolve the spanning controversy in the macro finance literature. We use the algorithm to construct macro-based bond return predictors in this study but SAGLasso should also be useful in similar big data applications in finance and economics. For instance, we may construct a real-time expectation factor using the SAGLasso algorithm and examine if the implied bond risk premia are consistent with those demanded by investors in history (Piazzesi et al. 2015). This would allow us to explore an alternative explanation for the spanning controversy: It is due to the discrepancy between the short-rate expectation of real-time investors and the ex post estimates of an econometrician (Cieslak 2018).[24] We may also expand the macro panel data to incorporate survey forecasts of macro variables, which are shown to provide additional information in term structure modeling (see, e.g., Chernov and Mueller 2012 and Kim and Orphanides 2012). We leave these questions to future research.

[24] In an earlier version of this paper, Huang and Shi (2010) provide evidence consistent with the potential mechanism suggested by Duffee (2011). As noted in Cieslak (2018), these different explanations of the spanning controversy are not, however, mutually exclusive because its resolution "depends on the particular variables that the econometrician assumes a part of his/her information set" (p. 3269).

References

Ang, A., and G. Bekaert. 2007. Stock return predictability: Is it there? *Review of Financial Studies* 20(3):651–707.

Bai, J., and S. Ng. 2008. Forecasting economic time series using targeted predictors. *Journal of Econometrics* 146(2):304–317.

Bair, E., T. Hastie, D. Paul, and R. Tibshirani. 2006. Prediction by supervised principal components. *Journal of the American Statistical Association* 101(473):119–137.

Bali, T. G., A. Goyal, D. Huang, F. Jiang, and Q. Wen. 2021. Different Strokes: Return Predictability Across Stocks and Bonds with Machine Learning and Big Data. *Working Paper, Georgetown University*.

Bauer, M. D., and J. D. Hamilton. 2018. Robust bond risk premia. *Review of Financial Studies* 31(2):399–448.

Bauer, M. D., and G. D. Rudebusch. 2016. Resolving the spanning puzzle in macro-finance term structure models. *Review of Finance* 21(2):511–553.

Berardi, A., M. Markovich, A. Plazzi, and A. Tamoni. 2021. Mind the (Convergence) Gap: Bond Predictability Strikes Back! *Management Science, Forthcoming*.

Bianchi, D., M. Büchner, and A. Tamoni. 2021. Bond Risk Premia with Machine Learning. *Review of Financial Studies* 34(2):1046–1089.

Campbell, J., and R. Shiller. 1991. Yield spreads and interest rate movements: A bird's eye view. *Review of Economic Studies* 58(3):495–514.

Campbell, J., and S. Thompson. 2008. Predicting excess stock returns out of sample: Can anything beat the historical average? *Review of Financial Studies* 21(4):1509–1531.

Chernov, M., and P. Mueller. 2012. The term structure of inflation expectations. *Journal of Financial Economics* 106(2):367–394.

Cieslak, A. 2018. Short-Rate Expectations and Unexpected Returns in Treasury Bonds. *Review of Financial Studies* 31(9):3265–3306.

Cieslak, A., and P. Povala. 2015. Expected returns in Treasury bonds. *Review of Financial Studies* 28(10):2859–2901.

Clark, T., and M. McCracken. 2001. Tests of equal forecast accuracy and encompassing for nested models. *Journal of Econometrics* 105(1):85–110.

Cochrane, J., and M. Piazzesi. 2005. Bond risk premia. *American Economic Review* 95(1):138–160.

Cooper, I., and R. Priestley. 2009. Time-varying risk premiums and the output gap. *Review of Financial Studies* 22(7):2801–2833.

Diebold, F. X., and R. S. Mariano. 1995. Comparing Predictive Accuracy. *Journal of Business & Economic Statistics* 13(3):253–263.

Duffee, G. R. 2002. Term premia and interest rate forecasts in affine models. *Journal of Finance* 57(1):405–443.

Duffee, G. R. 2010. Sharpe ratios in term structure models. *Working paper, Johns Hopkins University*.

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

Duffee, G. R. 2013a. Bond pricing and the macroeconomy. In G. M. Constantinides, M. Harris, and R. M. Stulz (Eds.), *Handbook of the Economics of Finance*, Volume 2B: Asset Pricing, pp. 907–968. North Holland.

Duffee, G. R. 2013b. Forecasting Interest Rates. In A. Timmermann and G. Elliott (Eds.), *Handbook of Economic Forecasting*, Volume 2A, pp. 385–426. Elsevier.

Duffie, D., and R. Kan. 1996. A yield-factor model of interest rates. *Mathematical Finance* 6:379–406.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani. 2004. Least angle regression. *Annals of Statistics*:407–451.

Ericsson, N. 1992. Parameter constancy, mean square forecast errors, and measuring forecast performance: An exposition, extensions, and illustration. *Journal of Policy Modeling* 14(4):465–495.

Fama, E., and R. Bliss. 1987. The information in long-maturity forward rates. *American Economic Review* 77:680–692.

Freyberger, J., A. Neuhierl, and M. Weber. 2020. Dissecting Characteristics Non-Parametrically. *Review of Financial Studies* 33(5):2326–2377.

Ghysels, E., C. Horan, and E. Moench. 2018. Forecasting through the rear-view mirror: Data revisions and bond return predictability. *Review of Financial Studies* 31(2):678–714.

Gibson, M. S., and M. Pritsker. 2000. Improving Grid-Based Methods for Estimating Value at Risk of Fixed Income Portfolios. *Journal of Risk* 3(Winter):65–89.

Goto, S., and Y. Xu. 2015. Improving mean variance optimization through sparse hedging restrictions. *Journal of Financial and Quantitative Analysis* 50(06):1415–1441.

Gu, S., B. Kelly, and D. Xiu. 2020. Empirical Asset Pricing via Machine Learning. *Review of Financial Studies* 33(5):2223–2273.

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

Hansen, L., and R. Hodrick. 1980. Forward exchange rates as optimal predictors of future spot rates: An econometric analysis. *Journal of Political Economy* 88(5):829–853.

Harvey, D., S. Leybourne, and P. Newbold. 1997. Testing the equality of prediction mean squared errors. *International Journal of Forecasting* 13(2):281–291.

He, X., G. Feng, J. Wang, and C. Wu. 2021. Predicting Individual Corporate Bond Returns. *Working Paper, City University of Hong Kong*.

Hodrick, R. 1992. Dividend yields and expected stock returns: Alternative procedures for inference and measurement. *Review of Financial Studies* 5(3):357–386.

Huang, J.-Z., R. Li, J. Ni, and Z. Shi. 2016. Forecasting bond returns using high-dimensional model selection. *MFA 2017 Chicago Meetings Paper*. Penn State University.

Huang, J.-Z., and Z. Shi. 2010. Determinants of Bond Risk Premia. *AFA 2011 Denver Meetings Paper*. Available at http://ssrn.com/paper=1573186.

Huang, J.-Z., Z. Shi, and W. Zhong. 2015. Model Selection for High-Dimensional Problems. In C.-F. Lee and J. C. Lee (Eds.), *Handbook of Financial Econometrics and Statistics*, Volume 4, Chapter 77, pp. 2093–2118. Springer-Verlag New York.

Ibragimov, R., and U. K. Müller. 2010. t-Statistic based correlation and heterogeneity robust inference. *Journal of Business & Economic Statistics* 28(4):453–468.

Jones, C. S., and S. Tuzel. 2013. New Orders and Asset Prices. *Review of Financial Studies* 26(1):115–157.

Joslin, S., A. Le, and K. J. Singleton. 2013. Why Gaussian macro-finance term structure models are (nearly) unconstrained factor-VARs. *Journal of Financial Economics* 109(3):604–622.

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

Joslin, S., K. J. Singleton, and H. Zhu. 2011. A new perspective on Gaussian dynamic term structure models. *Review of Financial Studies* 24(3):926–970.

Kim, D., and A. Orphanides. 2012. Term structure estimation with survey data on interest rate forecasts. *Journal of Financial and Quantitative Analysis* 47(1):241–272.

Le, A., and K. Singleton. 2013. The Structure of Risks in Equilibrium Affine Models of Bond Yields. *Working Paper, Kenan-Flagler Business School, UNC*.

Lewellen, J. 2015. The Cross-section of Expected Stock Returns. *Critical Finance Review* 4(1):1–44.

Litterman, R. B., and J. Scheinkman. 1991. Common factors affecting bond returns. *Journal of Fixed Income* 1(1):54–61.

Ludvigson, S., and S. Ng. 2009. Macro factors in bond risk premia. *Review of Financial Studies* 22(12):5027–5067.

Ludvigson, S., and S. Ng. 2011. A factor analysis of bond risk premia. In A. Ullah and D. E. A. Giles (Eds.), *Handbook of Empirical Economics and Finance*, pp. 313–372. CRC Press.

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

Piazzesi, M., , J. Salomao, and M. Schneider. 2015. Trend and cycle in bond premia. *Working Paper, Stanford University*.

Piazzesi, M., M. Schneider, and S. Tuzel. 2007. Housing, consumption and asset pricing. *Journal of Financial Economics* 83(3):531–569.

Stambaugh, R. 1988. The information in forward rates: Implications for models of the term structure. *Journal of Financial Economics* 21(1):41–70.

Stock, J., and M. Watson. 2002. Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association* 97(460):1167–1179.

Stock, J., and M. Watson. 2005. Implications of dynamic factor models for VAR analysis. *NBER working paper*.

Thornton, D. L., and G. Valente. 2012. Out-of-sample predictions of bond excess returns and forward rates: An asset allocation perspective. *Review of Financial Studies* 25(10):3141–3168.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B* 58(1):267–288.

Wachter, J. 2006. A consumption-based model of the term structure of interest rates. *Journal of Financial Economics* 79(2):365–399.

Wei, M., and J. H. Wright. 2013. Reverse Regressions And Long-Horizon Forecasting. *Journal of Applied Econometrics* 28(3):353–371.

Wright, J. H. 2011. Term Premia and Inflation Uncertainty: Empirical Evidence from an International Panel Dataset. *American Economic Review* 101(4):1514–1534.

Wright, J. H., and H. Zhou. 2009. Bond risk premia and realized jump risk. *Journal of Banking and Finance* 33(12):2333–2345.

Yuan, M., and Y. Lin. 2006. Model selection and estimation in regression with grouped variables. *Journal of Royal Statistical Society Series B* 68(1):49–67.

Zou, H. 2006. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* 101(476):1418–1429.

A Notation and Frequently Used Terms

Spanning hypothesis I (H_0^{S1})	Macro variables have no additional predictive power for excess bond returns over the first three principal components (PCs) of the observed yield curve
Spanning hypothesis II (H_0^{S2})	Macro variables have no additional predictive power for excess bond returns over the first five PCs of the noise-uncontaminated yield curve
Macro-unspanning hypothesis (H_0^{US})	So-called knife-edge restrictions given in Eq. (12) in the paper for a macro-finance term-structure model (MTSM) to be unspanned
$\widehat{CG}$	the convergence gap defined by Berardi et al. (2021)
$\widehat{CP}$	the Cochrane and Piazzesi (2005) forward rate factor
$\widehat{G}$	the (unconditional) the Supervised Adaptive Group LASSO (SAGLasso) macro factor constructed in this study
$\widetilde{G}$	the recursive SAGLasso macro factor constructed in this study
$\widehat{g}_1$, $\widehat{g}_2$, and $\widehat{g}_3$	(unconditional) SAGLasso group factors constructed in this study, representing "employment," "housing," and "inflation," respectively
$\widetilde{g}_1$, $\widetilde{g}_2$, and $\widetilde{g}_3$	recursively constructed $\widehat{g}_1$, $\widehat{g}_2$, and $\widehat{g}_3$
$\widehat{H}$	the hidden factor proposed by Duffee (2011)
$\widehat{LN}^m$	a modified Ludvigson and Ng (2009) macro-based return predictor
$PC^o_{1-3} = (PC^o_1, PC^o_2, PC^o_3)$	vector of the first three PCs of the *observed* yield curve
$\widetilde{PC}^o_{1-3}$	recursively constructed PC^o_{1-3}
$PC_{1-5} = (PC_1, \ldots, PC_5)$	vector of the first five PCs of the *noise-uncontaminated* yield curve
$\widetilde{PC}_{1-5}$	recursively constructed PC_{1-5}
$CSM(\mathcal{L},\mathcal{N})$	An $\mathcal{N}$-factor constrained, spanned MTSM—Model $SM(\mathcal{L},\mathcal{N})$ with restrictions on the model-implied Sharpe ratios of bond returns
$CUSM(\mathcal{L},\mathcal{N})$	An $\mathcal{N}$-factor constrained, unspanned MTSM—Model $USM(\mathcal{L},\mathcal{N})$ with restrictions on the model-implied Sharpe ratios of bond returns
$SM(\mathcal{L},\mathcal{N})$	An $\mathcal{N}$-factor spanned model—an $\mathcal{N}$-factor MTSM with $\mathcal{L}$ $(\mathcal{N}-1)$ yield factors and one macro factor (the SAGLasso factor G) that does not satisfy the macro-unspanning hypothesis H_0^{US}
$USM(\mathcal{L},\mathcal{N})$	An $\mathcal{N}$-factor spanned model—an $\mathcal{N}$-factor MTSM with $\mathcal{L}$ $(\mathcal{N}-1)$ yield factors and one macro factor (the SAGLasso factor G) that satisfies H_0^{US}
$YTSM(\mathcal{N})$	An $\mathcal{N}$-factor "yields-only" term-structure model

B Macroeconomic Series Used in the Analysis

Two sets of 131 macroeconomic series are used in our empirical analysis. The first, the standard one used in the literature, includes revised macroeconomic data. The second set consists of real-time macroeconomic data only—the macro series adjusted for data revisions and publication lags.

Table A.1 lists the 131 macroeconomic series and contains the full name (column 4) of each series, along with its series number (column 1), group number (column 2), mnemonic—the series label used in the source database (column 3), short name (column 5), and data transformation flag (column 6). The transformation *flag* = 1: no transformation applied to the series; *flag* = 2: the first difference applied; *flag* = 3: the second difference; *flag* = 4: the logarithm; *flag* = 5: the first difference of logarithm; and *flag* = 6: the second difference of logarithm.[25]

We compile our macro data in three steps. First, we match the panel of 131 series with ALFRED and find that 70 of them are included in the latter. For each of the 70 matched series, we collect its latest *nine* real-time observations at the end of each month (we do this because some macro variables need to be transformed to their second-order log-differences). However, vintage versions of these 70 series are not balanced and go back to 1964 for only 25 series. Nonetheless, only 3 out of the 19 macro variables eventually selected by SAGLasso do *not* have their vintage data available going back to January 1985. Therefore, the look-forward biases should have a minimum impact, at least on our results obtained from the post-1984 sample.

Second, for the 45 incomplete series in ALFRED, we fill in their missing observations using data over 1964–2007 provided by Ludvigson and Ng (2011) and our manually updated observations from the Federal Reserve Economic Data and The Conference Board over the post-2007 period. As for the 61 series not included in ALFRED, these variables are presumably not subject to revision.[26] We obtain observations for these 61 series from the aforementioned two sources. We then adjust all these macro variables for their publication lags; that is, for each of these time series, we calculate

[25] Second-order log-differences are the reason for keeping the latest nine observations at each point of historical time for each of the 70 matched series in ALFRED (see Section 3). To see that, let $x_{s|t}$ denote the value of a particular macro variable collected for calendar month s at the end of month $t \geq s$. Suppose that this variable is released with a one-month lag and needs to be log-differenced twice to attain stationarity. The final data to be included in the SAGLasso procedures would be $\{\Delta^2 \ln x_{t-1|t}, \Delta^2 \ln x_{t-2|t}, \ldots, \Delta^2 \ln x_{t-8|t}\}$, where $\Delta^2 \ln x_{t-1|t} = \ln x_{t-1|t} - 2 \ln x_{t-2|t} + \ln x_{t-3|t}$.

[26] This conjecture is partially confirmed by checking observations of these macro series around the end of 2007. The logic is as follows. The LN09 data set ceases its coverage of macro time series in December 2007. If a specific macroeconomic measure (not included in ALFRED) is subject to data revision, its observations for the last couple of months in their data set are likely from the first (preliminary) and second releases. These observations are then compared with corresponding ones collected in 2015, which are definitely from the third (final) release. We find that they are identical. Regardless, the main findings of this study are not affected by this conjecture. As mentioned earlier, it turns out that among those macro series included in the SAGLasso factor, only three commodity price indices have no vintage data available, and these indices should not be subject to revision.

the integer number of months in the time interval between the end of the period over which it is measured and its release date. As shown later, such adjustments matter in our predictability analysis.

Finally, we investigate the time-series properties of these 131 series and determine transformations needed to stationarize each of these series. Table A.1 provides a complete list of the 131 series and, for each series, its data transforms applied, its publication lag, and the availability of its vintage data.

Column 7 labeled "$\widehat{G}_t$" of Table A.1 shows the values of a flag indicating which of the 131 macroeconomic series has a nonzero coefficient for its contemporaneous and/or lagged values (up to 6) in the SAGLasso regression. The flag value of "0" corresponds to the contemporaneous variable, and the value of "ℓ" denotes lag ℓ (in months), $\ell = 1, \ldots, 6$. For instance, macro series #41 (CES048) in group 2—which measures the employment situation in the industry sector "Trade, Transportation and Utilities"—is selected by the SAGLasso approach and has 2 variables (out of 7), the lag-5 and lag-6 values of the series, included in the SAGLasso macro factor $\widehat{G}$. In total, 19 out of the 131 series (30 out of the 917 macro variables) enter the $\widehat{G}$ factor. Column 9 labeled "Lag" reports each series' publication lag (in months), which is defined as the time between the end of the period over which the series is measured and its first release date. Note that out of the 131 series, the four in group 8 "stock market" (#81 through #84) are the only ones without a publication delay. The last column, labeled "vintage," indicates which macro series has vintage data available, where an asterisk denotes those series whose real-time series are available and used in our empirical analysis. Note that out of the 19 series included in the $\widehat{G}$ factor and two additional series (#42 and #53) included in $\tilde{G}$ (the out-of-sample version of $\widehat{G}$), the three commodity price indices (#111 through #113) are the only series that have no vintage data available in the ALFRED database. However, given the nature of these three series, they should not be subject to revision.

C Supervised Adaptive Group Lasso Method

We first briefly review the group lasso (Yuan and Lin 2006). We begin with the following model:

$$\mathbf{Y} = \mathbf{X}\beta^{\mathbf{0}} + e, \tag{15}$$

where e is assumed to be a T-dimensional vector of *i.i.d.* errors (we will relax this assumption later). The main assumption of the Group Lasso is that some subvectors of the true coefficients $\beta^{\mathbf{0}}$ are zero. We denote by $h \in \mathcal{H}_1 = \{h : \beta^{\mathbf{0}}_{\mathbf{h}} \neq 0\}$ the unknown index set of non-zero subvectors of

$\beta^{\mathbf{0}}$. Hence, the Group Lasso involves identifying $\mathcal{H}_1$ and estimating $\beta^{\mathbf{0}}$.

The method is usually implemented by estimating the following restrictive form:

$$\min_{\beta \in \mathbb{R}^N} \left\{ \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda \sum_h \|\beta_h\| \right\}. \tag{16}$$

Note that expression (16) reduces to the Lasso when $|\mathcal{H}| = N$ and each h corresponds to the one-dimensional subspace of $\mathbb{R}^T$ spanned by the corresponding column of the design matrix $\mathbf{X}$. In our implementation, we consider the general Group Lasso and more specifically, the adaptive group lasso, as follows:

$$\min_{\beta \in \mathbb{R}^N} \left\{ \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda \sum_h w_h \|\beta_h\| \right\}. \tag{17}$$

Next, we describe the Supervised Adaptive Group Lasso (SAGLasso) algorithm proposed in Section 4.1. The method consists of two steps.

Step I: For cluster $h \in \mathcal{H}$, compute $\hat{\beta}^h$—the cluster-wise Adaptive Lasso estimate of β^h, namely,

$$\hat{\beta}^h = \underset{\beta^h}{\operatorname{argmin}} \left\{ \|\mathbf{arx} - \mathbf{X}_h \beta^h\|^2 + \sum_j \lambda_h * \hat{w}_{hj} |\beta_j^h| \right\}, \tag{18}$$

where $\mathbf{arx}$ is a vector of average excess bond returns across maturity and $\hat{w}_{hj}$ the j-th component of $\hat{\mathbf{w}}_h$, the vector of the (adaptive) weights. Zou (2006) recommends using $\hat{\beta}^{OLS}$ to construct $\hat{\mathbf{w}}_h$. As collinearity is a concern in our case, we set $\hat{\mathbf{w}}_h = 1/|\hat{\beta}_h^{RID}|^{\gamma_h}$, where $\hat{\beta}_h^{RID}$ is the best ridge regression fit of $\mathbf{arx}$ on $\mathbf{X}_h$. That is, for cluster h we only use macroeconomic variables within that cluster to construct predictive models. The optimal pairs of (γ_h, λ_h) are determined using five-dimensional cross-validations. It is worth noting that tuning parameters λ_h are selected for each cluster separately in order to have different degrees of regularization for different clusters. This flexibility allows us to uncover subtle structures that otherwise will be missed when applying the (adaptive) lasso method to all the series/clusters at the same time.

Note that for each cluster $h \in \mathcal{H}$, the adaptive lasso $\hat{\beta}^h$ has only a small number of nonzero components. Let $\tilde{\beta}^h = \hat{\beta}^h \setminus \mathbf{0}$, the vector of nonzero estimated components of $\hat{\beta}^h$ given by the cluster-wise model (18), and denote the corresponding part of $\mathbf{X}_h$ by $\tilde{X}_h$. In our case, a typical cluster size ($\dim(\mathbf{X}_h)$) of 80 variables may reduce to a $\dim(\tilde{X}_h)$ of $8 \sim 10$. Namely, the number of macro variables selected in Step I is significantly smaller than the original number to begin with.

Step II: Construct the joint predictive model under the Group Lasso constraint as follows:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \|\mathbf{arx} - \tilde{\mathbf{X}}\beta\|^2 + \lambda \sum_{h \in \mathcal{H}} w_h \|\beta_h\| \right\}, \tag{19}$$

where $\tilde{\mathbf{X}}$ is formed by concatenating the design matrices $\tilde{X}_h$. The parameter λ is also chosen by five-fold cross-validation. With $\lambda \to \infty$, estimates of some components of $\tilde{\beta}_h$s can be exactly zero. Following Yuan and Lin (2006), we obtain the solution in Eq. (19) efficiently by using the modified least angle regression selection algorithm of Efron et al. (2004).

In out-of-sample tests conducted in our analysis, tuning parameters $\{\lambda_h, \lambda\}$ are selected recursively starting from the beginning of the test period using cross-validation as well as information only available at the time of estimation. However, to reduce the bias due to the limited training sample size, we use ten-fold cross-validation for the first five years of the out-of-sample testing period (e.g., the period 1985–1989 for the full sample). After that we go back to standard five-fold cross-validation to restore the balance between bias and variance. Also, to reduce the computational burden in the finite-sample analysis (Section 5.2.2), we select $\{\lambda_h, \lambda\}$ once for each quarter rather than for each month; that is, $\{\lambda_h, \lambda\}$ selected in January are also used to perform SAGLasso model selection in February and March, until they are reselected in April.

Note that the SAGLasso algorithm differs from the supervised principal component analysis (SPCA)—another two-step supervised learning approach—proposed by Bair et al. (2006) in a biological setting, which has been applied to inflation forecasts in Bai and Ng (2008).[27] For instance, the former takes into account the underlying cluster structure of candidate variables, whereas the SPCA does not consider all the candidates simultaneously. Also, variables selected in the SPCA are the PCs whose economic interpretations may not be obvious even though they may have satisfactory prediction performance. Factors constructed using SAGLasso, however, are easier to interpret.

Group Lasso is also applied by Freyberger et al. (2020) to identify firm characteristics in shaping expected equity returns. In their analysis, each group consists of 20 portfolios associated with (a polynomial function of) one characteristic, and model selection is done at the group level only. In our analysis, each group consists of macro variables supposed to capture the same economic concept, and Adaptive Lasso is used within each group (before model selection at the group level) to further mitigate the curse of dimensionality and boost the out-of-sample performance.

[27] Gibson and Pritsker (2000) use partial least squares to choose risk factors of fixed-income portfolios. Goto and Xu (2015) apply the graphical lasso to portfolio selection.

Table A.1: Macro Data Description

Series No.	Group	Mnemonic	Description	Short Name	tran	$\widehat{G}_t$	Lag	Vintage
1	1	a0m052	Personal income (AR, bil. chain 2000 $)	PI	5		1	*
2	1	A0M051	Personal income less transfer payments (AR, bil. chain 2000 $)	PI less transfers	5		1	*
3	4	A0M224R	Real Consumption (AC) A0m224/gmdc	Consumption	5		1	*
4	4	A0M057	Manufacturing and trade sales (mil. Chain 1996 $)	M & T sales	5		1	
5	4	A0M059	Sales of retail stores (mil. Chain 2000 $)	Retail sales	5		1	
6	1	IPS10	INDUSTRIAL PRODUCTION INDEX - TOTAL INDEX	IP: total	5		1	*
7	1	IPS11	INDUSTRIAL PRODUCTION INDEX - PRODUCTS, TOTAL	IP: products	5		1	*
8	1	IPS299	INDUSTRIAL PRODUCTION INDEX - FINAL PRODUCTS	IP: final prod	5		1	*
9	1	IPS12	INDUSTRIAL PRODUCTION INDEX - CONSUMER GOODS	IP: cons gds	5		1	*
10	1	IPS13	INDUSTRIAL PRODUCTION INDEX - DURABLE CONSUMER GOODS	IP: cons dble	5		1	*
11	1	IPS18	INDUSTRIAL PRODUCTION INDEX - NONDURABLE CONSUMER GOODS	iIP:cons nondble	5		1	*
12	1	IPS25	INDUSTRIAL PRODUCTION INDEX - BUSINESS EQUIPMENT	IP:bus eqpt	5		1	*
13	1	IPS32	INDUSTRIAL PRODUCTION INDEX - MATERIALS	IP: matls	5		1	*
14	1	IPS34	INDUSTRIAL PRODUCTION INDEX - DURABLE GOODS MATERIALS	IP: dble mats	5		1	*
15	1	IPS38	INDUSTRIAL PRODUCTION INDEX - NONDURABLE GOODS MATERIALS	IP:nondble mats	5		1	*
16	1	IPS43	INDUSTRIAL PRODUCTION INDEX - MANUFACTURING (SIC)	IP: mfg	5		1	*
17	1	IPS307	INDUSTRIAL PRODUCTION INDEX - RESIDENTIAL UTILITIES	IP: res util	5		1	*
18	1	IPS306	INDUSTRIAL PRODUCTION INDEX - FUELS	IP: fuels	5		1	*
19	1	PMP	NAPM PRODUCTION INDEX (PERCENT)	NAPM prodn	1		1	
20	1	A0m082	Capacity Utilization (Mfg)	Cap util	2		1	
21	2	LHEL	INDEX OF HELP-WANTED ADVERTISING IN NEWSPAPERS (1967=100;SA)	Help wanted indx	2		1	
22	2	LHELX	EMPLOYMENT: RATIO; HELP-WANTED ADS:NO. UNEMPLOYED CLF	Help wanted/emp	2		1	
23	2	LHEM	CIVILIAN LABOR FORCE: EMPLOYED, TOTAL (THOUS.,SA)	Emp CPS total	5		1	*
24	2	LHNAG	CIVILIAN LABOR FORCE: EMPLOYED, NONAGRIC.INDUSTRIES (THOUS.,SA)	Emp CPS nonag	5		1	*
25	2	LHUR	UNEMPLOYMENT RATE: ALL WORKERS, 16 YEARS & OVER (%,SA)	U: all	2		1	*
26	2	LHU680	UNEMPLOY.BY DURATION: AVERAGE(MEAN)DURATION IN WEEKS (SA)	U: mean duration	2		1	*
27	2	LHU5	UNEMPLOY.BY DURATION: PERSONS UNEMPL.LESS THAN 5 WKS (THOUS.,SA)	U $<$ 5 wks	5		1	
28	2	LHU14	UNEMPLOY.BY DURATION: PERSONS UNEMPL.5 TO 14 WKS (THOUS.,SA)	U 5-14 wks	5		1	
29	2	LHU15	UNEMPLOY.BY DURATION: PERSONS UNEMPL.15 WKS + (THOUS.,SA)	U 15+ wks	5		1	
30	2	LHU26	UNEMPLOY.BY DURATION: PERSONS UNEMPL.15 TO 26 WKS (THOUS.,SA)	U 15-26 wks	5		1	
31	2	LHU27	UNEMPLOY.BY DURATION: PERSONS UNEMPL.27 WKS + (THOUS,SA)	U 27+ wks	5		1	
32	2	A0M005	Average weekly initial claims, unemploy. insurance (thous.)	UI claims	5		1	*
33	2	CES002	EMPLOYEES ON NONFARM PAYROLLS - TOTAL PRIVATE	Emp: total	5		1	*
34	2	CES003	EMPLOYEES ON NONFARM PAYROLLS - GOODS-PRODUCING	Emp: gds prod	5		1	*
35	2	CES006	EMPLOYEES ON NONFARM PAYROLLS - MINING	Emp: mining	5		1	*
36	2	CES011	EMPLOYEES ON NONFARM PAYROLLS - CONSTRUCTION	Emp: const	5		1	*
37	2	CES015	EMPLOYEES ON NONFARM PAYROLLS - MANUFACTURING	Emp: mfg	5		1	*
38	2	CES017	EMPLOYEES ON NONFARM PAYROLLS - DURABLE GOODS	Emp: dble gds	5		1	*
39	2	CES033	EMPLOYEES ON NONFARM PAYROLLS - NONDURABLE GOODS	Emp: nondbles	5		1	*
40	2	CES046	EMPLOYEES ON NONFARM PAYROLLS - SERVICE-PROVIDING	Emp: services	5		1	*
41	2	CES048	EMPLOYEES ON NONFARM PAYROLLS - TRADE, TRANSPORTATION, AND UTILITIES	Emp: TTU	5	5,6	1	*
42	2	CES049	EMPLOYEES ON NONFARM PAYROLLS - WHOLESALE TRADE	Emp: wholesale	5		1	*
43	2	CES053	EMPLOYEES ON NONFARM PAYROLLS - RETAIL TRADE	Emp: retail	5		1	*
44	2	CES088	EMPLOYEES ON NONFARM PAYROLLS - FINANCIAL ACTIVITIES	Emp: FIRE	5	0,1,2,3	1	*
45	2	CES140	EMPLOYEES ON NONFARM PAYROLLS - GOVERNMENT	Emp: Govt	5		1	*
46	2	CES151	AVERAGE WEEKLY HOURS OF PRODUCTION OR NONSUPERVISORY WORKERS ON PRIVATE NONFAR	Avg hrs	1	0,2	1	*
47	2	CES155	AVERAGE WEEKLY HOURS OF PRODUCTION OR NONSUPERVISORY WORKERS ON PRIVATE NONFAR	Overtime: mfg	2		1	*
48	2	aom001	Average weekly hours, mfg. (hours)	Avg hrs: mfg	1	0,2	1	*
49	2	PMEMP	NAPM EMPLOYMENT INDEX (PERCENT)	NAPM empl	1	0	1	*
50	3	HSFR	HOUSING STARTS:NONFARM(1947-58);TOTAL FARM & NONFARM(1959-)(THOUS.,SA	HStarts: Total	5	5	1	*
51	3	HSNE	HOUSING STARTS:NORTHEAST (THOUS.U.)S.A.	HStarts: NE	4	2,6	1	*
52	3	HSMW	HOUSING STARTS:MIDWEST(THOUS.U.)S.A.	HStarts: MW	4	2	1	*
53	3	HSSOU	HOUSING STARTS:SOUTH (THOUS.U.)S.A.	HStarts: South	4		1	*
54	3	HSWST	HOUSING STARTS:WEST (THOUS.U.)S.A.	HStarts: West	4	1,3	1	*
55	3	HSBR	HOUSING AUTHORIZED: TOTAL NEW PRIV HOUSING UNITS (THOUS.,SAAR)	BP: total	4	3,6	1	*
56	3	HSBNE	HOUSES AUTHORIZED BY BUILD. PERMITS:NORTHEAST(THOU.U.)S.A	BP: NE	4	0	1	*
57	3	HSBMW	HOUSES AUTHORIZED BY BUILD. PERMITS:MIDWEST(THOU.U.)S.A.	BP: MW	4		1	*
58	3	HSBSOU	HOUSES AUTHORIZED BY BUILD. PERMITS:SOUTH(THOU.U.)S.A.	BP: South	4	0,6	1	*
59	3	HSBWST	HOUSES AUTHORIZED BY BUILD. PERMITS:WEST(THOU.U.)S.A.	BP: West	4	3,6	1	*
60	4	PMI	PURCHASING MANAGERS' INDEX (SA)	PMI	1		1	
61	4	PMNO	NAPM NEW ORDERS INDEX (PERCENT)	NAPM new ordrs	1		2	
62	4	PMDEL	NAPM VENDOR DELIVERIES INDEX (PERCENT)	NAPM vendor del	1		2	
63	4	PMNV	NAPM INVENTORIES INDEX (PERCENT)	NAPM Invent	1		2	
64	4	A0M008	Mfrs' new orders, consumer goods and materials (bil. chain 1982 $)	Orders: cons gds	5		2	
65	4	A0M007	Mfrs' new orders, durable goods industries (bil. chain 2000 $)	Orders: dble gds	5		2	
66	4	A0M027	Mfrs' new orders, nondefense capital goods (mil. chain 1982 $)	Orders: cap gds	5		2	

Series No.	Group	Mnemonic	Description	Short Name	tran	$\widehat{G}_t$	Lag	Vintage
67	4	A1M092	Mfrs' unfilled orders, durable goods indus. (bil. chain 2000 $)	Unf orders: dble	5		1	
68	4	A0M070	Manufacturing and trade inventories (bil. chain 2000 $)	M & T invent	5		2	
69	4	A0M077	Ratio, mfg. and trade inventories to sales (based on chain 2000 $)	M & T invent/sales	2		2	
70	5	FM1	MONEY STOCK: M1(CURR,TRAV.CKS,DEM DEP,OTHER CK'ABLE DEP)(BIL$,SA)	M1	6		1	*
71	5	FM2	MONEY STOCK:M2(M1+O'NITE RPS,EURO$,G/P&B/D MMMFS&SAV&SM TIME DEP(BIL$,	M2	6		1	*
72	5	FM3	MONEY STOCK: M3(M2+LG TIME DEP,TERM RP'S&INST ONLY MMMFS)(BIL$,SA)	M3	6		1	*
73	5	FM2DQ	MONEY SUPPLY - M2 IN 1996 DOLLARS (BCI)	M2 (real)	5		1	*
74	5	FMFBA	MONETARY BASE, ADJ FOR RESERVE REQUIREMENT CHANGES(MIL$,SA)	MB	6		1	
75	5	FMRRA	DEPOSITORY INST RESERVES:TOTAL,ADJ FOR RESERVE REQ CHGS(MIL$,SA)	Reserves tot	6		1	
76	5	FMRNBA	DEPOSITORY INST RESERVES:NONBORROWED,ADJ RES REQ CHGS(MIL$,SA)	Reserves nonbor	6		1	
77	5	FCLNQ	COMMERCIAL & INDUSTRIAL LOANS OUSTANDING IN 1996 DOLLARS (BCI)	C&I loans	6		1	
78	5	FCLBMC	WKLY RP LG COM'L BANKS:NET CHANGE COM'L & INDUS LOANS(BIL$,SAAR)	C&I loans	1		1	
79	5	CCINRV	CONSUMER CREDIT OUTSTANDING - NONREVOLVING(G19)	Cons credit-Nonrevolving	6		1	
80	5	A0M095	Ratio, consumer installment credit to personal income (pct.)	Inst cred/PI	2		1	
81	8	FSPCOM	S&P'S COMMON STOCK PRICE INDEX: COMPOSITE (1941-43=10)	S&P 500	5		0	
82	8	FSPIN	S&P'S COMMON STOCK PRICE INDEX: INDUSTRIALS (1941-43=10)	S&P: indust	5		0	
83	8	FSDXP	S&P'S COMPOSITE COMMON STOCK: DIVIDEND YIELD (% PER ANNUM)	S&P div yield	2		0	
84	8	FSPXE	S&P'S COMPOSITE COMMON STOCK: PRICE-EARNINGS RATIO (%,NSA)	S&P PE ratio	5		0	
85	6	FYFF	INTEREST RATE: FEDERAL FUNDS (EFFECTIVE) (% PER ANNUM,NSA)	FedFunds	2		1	
86	6	CP90	Cmmercial Paper Rate (AC)	Commpaper	2		1	
87	6	FYGM3	INTEREST RATE: U.S.TREASURY BILLS,SEC MKT,3-MO.(% PER ANN,NSA)	3 mo T-bill	2		1	
88	6	FYGM6	INTEREST RATE: U.S.TREASURY BILLS,SEC MKT,6-MO.(% PER ANN,NSA)	6 mo T-bill	2		1	
89	6	FYGT1	INTEREST RATE: U.S.TREASURY CONST MATURITIES,1-YR.(% PER ANN,NSA)	1 yr T-bond	2		1	
90	6	FYGT5	INTEREST RATE: U.S.TREASURY CONST MATURITIES,5-YR.(% PER ANN,NSA)	5 yr T-bond	2		1	
91	6	FYGT10	INTEREST RATE: U.S.TREASURY CONST MATURITIES,10-YR.(% PER ANN,NSA)	10 yr T-bond	2		1	
92	6	FYAAAC	BOND YIELD: MOODY'S AAA CORPORATE (% PER ANNUM)	Aaabond	2		1	
93	6	FYBAAC	BOND YIELD: MOODY'S BAA CORPORATE (% PER ANNUM)	Baa bond	2		1	
94	6	scp90	cp90-fyff	CP-FF spread	1		1	
95	6	sfygm3	fygm3-fyff	3 mo-FF spread	1		1	
96	6	sFYGM6	fygm6-fyff	6 mo-FF spread	1		1	
97	6	sFYGT1	fygt1-fyff	1 yr-FF spread	1		1	
98	6	sFYGT5	fygt5-fyff	5 yr-FFspread	1		1	
99	6	sFYGT10	fygt10-fyff	10yr-FF spread	1		1	
100	6	sFYAAAC	fyaaac-fyff	Aaa-FF spread	1		1	
101	6	sFYBAAC	fybaac-fyff	Baa-FF spread	1		1	
102	6	EXRUS	UNITED STATES;EFFECTIVE EXCHANGE RATE(MERM)(INDEX NO.)	Ex rate: avg	5		2	
103	6	EXRSW	FOREIGN EXCHANGE RATE: SWITZERLAND (SWISS FRANC PER U.S.$)	Ex rate: Switz	5		1	
104	6	EXRJAN	FOREIGN EXCHANGE RATE: JAPAN (YEN PER U.S.$)	Ex rate: Japan	5		1	
105	6	EXRUK	FOREIGN EXCHANGE RATE: UNITED KINGDOM (CENTS PER POUND)	Ex rate: UK	5		1	
106	6	EXRCAN	FOREIGN EXCHANGE RATE: CANADA (CANADIAN *PERU.S.*)	EX rate: Canada	5		1	
107	7	PWFSA	PRODUCER PRICE INDEX: FINISHED GOODS (82=100,SA)	PPI: fin gds	6		1	*
108	7	PWFCSA	PRODUCER PRICE INDEX:FINISHED CONSUMER GOODS (82=100,SA)	PPI: cons gds	6		1	*
109	7	PWIMSA	PRODUCER PRICE INDEX:INTERMED MAT.SUPPLIES & COMPONENTS(82=100,SA)	PPI: int matls	6		1	*
110	7	PWCMSA	PRODUCER PRICE INDEX:CRUDE MATERIALS (82=100,SA)	PPI: crude matls	6		1	*
111	7	PSCCOM	SPOT MARKET PRICE INDEX:BLS & CRB: ALL COMMODITIES(1967=100)	Commod: spot price	6	0	1	
112	7	PSM99Q	INDEX OF SENSITIVE MATERIALS PRICES (1990=100)(BCI-99A)	Sens matls price	6	0	1	
113	7	PMCP	NAPM COMMODITY PRICES INDEX (PERCENT)	NAPM com price	1	0	1	
114	7	PUNEW	CPI-U: ALL ITEMS (82-84=100,SA)	CPI-U: all	6		1	*
115	7	PU83	CPI-U: APPAREL & UPKEEP (82-84=100,SA)	CPI-U: apparel	6		1	*
116	7	PU84	CPI-U: TRANSPORTATION (82-84=100,SA)	CPI-U: transp	6		1	*
117	7	PU85	CPI-U: MEDICAL CARE (82-84=100,SA)	CPI-U: medical	6		1	*
118	7	PUC	CPI-U: COMMODITIES (82-84=100,SA)	CPI-U: comm.	6		1	*
119	7	PUCD	CPI-U: DURABLES (82-84=100,SA)	CPI-U: dbles	6		1	*
120	7	PUS	CPI-U: SERVICES (82-84=100,SA)	CPI-U: services	6		1	*
121	7	PUXF	CPI-U: ALL ITEMS LESS FOOD (82-84=100,SA)	CPI-U: ex food	6		1	*
122	7	PUXHS	CPI-U: ALL ITEMS LESS SHELTER (82-84=100,SA)	CPI-U: ex shelter	6	6	1	*
123	7	PUXM	CPI-U: ALL ITEMS LESS MEDICAL CARE (82-84=100,SA)	CPI-U: ex med	6		1	*
124	7	GMDC	PCE,IMPL PR DEFL:PCE (1987=100)	PCE defl	6		2	*
125	7	GMDCD	PCE,IMPL PR DEFL:PCE; DURABLES (1987=100)	PCE defl: dlbes	6	4	2	*
126	7	GMDCN	PCE,IMPL PR DEFL:PCE; NONDURABLES (1996=100)	PCE defl: nondble	6		2	*
127	7	GMDCS	PCE,IMPL PR DEFL:PCE; SERVICES (1987=100)	PCE defl: services	6	6	2	*
128	2	CES275	AVERAGE HOURLY EARNINGS OF PRODUCTION OR NONSUPERVISORY WORKERS ON PRIVATE NO	AHE: goods	6		1	
129	2	CES277	AVERAGE HOURLY EARNINGS OF PRODUCTION OR NONSUPERVISORY WORKERS ON PRIVATE NO	AHE: const	6		1	
130	2	CES278	AVERAGE HOURLY EARNINGS OF PRODUCTION OR NONSUPERVISORY WORKERS ON PRIVATE NO	AHE: mfg	6		1	
131	4	HHSNTN	U. OF MICH. INDEX OF CONSUMER EXPECTATIONS(BCD-83)	Consumer expect	2		1	

Table 1: Unconditional and Conditional Predictive Power of the SAGLasso Macro Factor

The return to an n-year zero-coupon bond from month t to month $t+12$ less the month-t yield on a one-year bond is regressed on the SAGLasso macro factor $\widehat{G}_t$ alone (columns 1-4) or $\widehat{G}_t$ and another return predictor together. The latter predictor considered includes the modified Ludvigson and Ng (2011) factor $\widehat{LN}_t^m$ (columns 5-8); the Cochrane and Piazzesi (2005) forward-rate factor $\widehat{CP}_t$ (columns 9-12); the hidden component $\widehat{H}_t$ in the one-dimensional term premium factor as constructed in Duffee (2011) (columns 13-16); and the Convergence Gap $\widehat{CG}_t$ as defined in Berardi et al. (2021) (columns 17-20). Test statistics are computed using the Hansen and Hodrick (1980) GMM covariance estimator (in parentheses), or the Newey and West (1987) HAC covariance estimator (in brackets). Bond maturity n equals 2, 3, 4, and 5 for the full sample spanning the period January 1964–December 2014 (Panel A), and equals 2, 5, 7, and 10 for the post-1984 sample period January 1985–December 2014 (Panel B).

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)
Panel A: Full sample, 1964–2014																				
maturity n (year)	2	3	4	5	2	3	4	5	2	3	4	5	2	3	4	5	2	3	4	5
$\widehat{G}_t$	1.047	1.982	2.810	3.498	0.924	1.671	2.301	2.807	0.889	1.677	2.314	2.929	1.038	1.962	2.783	3.471	0.924	1.836	2.667	3.368
	(5.911)	(6.431)	(6.753)	(7.003)	(4.704)	(4.960)	(4.980)	(5.013)	(5.039)	(5.737)	(6.276)	(6.466)	(6.679)	(7.715)	(7.840)	(7.825)	(4.652)	(5.044)	(5.277)	(5.557)
	[6.427]	[6.954]	[7.323]	[7.607]	[5.206]	[5.483]	[5.542]	[5.595]	[5.477]	[6.172]	[6.736]	[6.989]	[7.277]	[8.357]	[8.550]	[8.546]	[5.152]	[5.589]	[5.866]	[6.184]
$\widehat{LN}_t^m$					0.071	0.179	0.293	0.398												
					(1.451)	(2.076)	(2.609)	(3.026)												
					[1.601]	[2.300]	[2.904]	[3.355]												
$\widehat{CP}_t$									0.235	0.453	0.737	0.845								
									(2.850)	(2.818)	(3.246)	(2.972)								
									[3.187]	[3.186]	[3.681]	[3.361]								
$\widehat{H}_t$													−0.234	−0.531	−0.713	−0.713				
													(−2.363)	(−3.255)	(−3.159)	(−2.515)				
													[−2.464]	[−3.336]	[−3.250]	[−2.599]				
$\widehat{CG}_t$																	−0.172	−0.203	−0.199	−0.181
																	(−2.681)	(−1.635)	(−1.133)	(−0.861)
																	[−2.835]	[−1.743]	[−1.218]	[−0.921]
R^2	0.352	0.375	0.388	0.392	0.365	0.398	0.420	0.431	0.398	0.425	0.457	0.451	0.396	0.441	0.449	0.432	0.401	0.395	0.398	0.397
Panel B: Subsample, 1985-2014																				
maturity n (year)	2	5	7	10	2	5	7	10	2	5	7	10	2	5	7	10	2	5	7	10
$\widehat{G}_t$	0.729	2.900	4.096	5.225	0.706	2.634	3.740	4.743	0.651	2.611	3.657	4.536	0.735	2.894	4.011	5.150	0.676	2.794	4.006	5.147
	(5.278)	(7.584)	(6.297)	(5.680)	(5.550)	(6.027)	(4.991)	(4.574)	(4.698)	(6.234)	(5.731)	(5.252)	(5.346)	(7.569)	(6.677)	(5.979)	(5.278)	(7.062)	(5.781)	(5.327)
	[5.731]	[7.945]	[6.585]	[5.989]	[5.876]	[6.392]	[5.288]	[4.857]	[5.085]	[6.717]	[6.118]	[5.638]	[5.813]	[7.942]	[6.958]	[6.298]	[5.605]	[7.379]	[6.085]	[5.642]
$\widehat{LN}_t^m$					0.030	0.333	0.447	0.604												
					(0.426)	(1.916)	(1.801)	(1.945)												
					[0.462]	[2.107]	[1.988]	[2.148]												
$\widehat{CP}_t$									0.149	0.551	0.837	1.315								
									(1.405)	(1.738)	(1.993)	(2.708)								
									[1.583]	[1.953]	[2.194]	[2.953]								
$\widehat{H}_t$													−0.001	0.002	0.021	0.019				
													(−0.347)	(0.121)	(1.241)	(1.065)				
													[−0.378]	[0.131]	[1.345]	[1.139]				
$\widehat{CG}_t$																	−0.340	−0.675	−0.575	−0.500
																	(−3.612)	(−2.548)	(−1.344)	(−0.894)
																	[−3.979]	[−2.749]	[−1.442]	[−0.963]
R^2	0.301	0.397	0.393	0.398	0.303	0.416	0.410	0.416	0.331	0.431	0.432	0.457	0.303	0.397	0.410	0.406	0.413	0.434	0.406	0.404

Table 2: Out-of-Sample Performance Assessment

Panel A reports accuracy of out-of-sample forecasts from models with and without the real-time macro factor $\tilde{G}$ as a return predictor. Benchmark predictors considered include the first three principal components (PCs) of observed yields ($\widetilde{PC}^{o}_{1-3,t}$) and the first five PCs of the noise-uncontaminated yield curve ($\widetilde{PC}_{1-5,t}$). The rows labeled "ENC-REG" report the out-of-sample t-statistics proposed by Ericsson (1992), and those labeled "ENC-NEW" report a variant of the ENC-REG statistic proposed by Clark and McCracken (2001); both tests share the same null hypothesis that the benchmark model encompasses the unrestricted model with excess predictors. "R^2_{oos}" denotes the out-of-sample R^2 of Campbell and Thompson (2008), and the rows labeled "ΔR^2_{oos}" represent the incremental R^2_{oos} due to $\tilde{G}$. Panel B reports the certainty equivalent gains (in percentage) for a mean-variance investor who selects an N-year bond ($N \geq 2$) along with a 1-year bond and who uses portfolios weights potentially depending on $\tilde{G}$-based forecasts. The investor's risk aversion coefficient γ is assumed to be either 3 or 5. The p-values of certainty equivalent returns (in angle brackets) are based on an extended version of Diebold and Mariano (1995) test. All out-of-sample forecasts are formed recursively, with a "training" period of 20 years for the entire sample or that of 15 years in the subsample analysis.

	Full sample, 1964–2014				Subsample, 1985–2014			
maturity (year)	2	3	4	5	2	5	7	10
	Panel A: Statistical significance							
R^2_{oos}	0.123	0.187	0.226	0.246	0.033	0.248	0.236	0.205
	Panel A1: $\tilde{G}_t$+$\widetilde{PC}^{o}_{1-3,t}$ vs. $\widetilde{PC}^{o}_{1-3,t}$							
ENC-REG	4.764	4.987	4.831	4.871	3.539	4.570	4.804	5.258
ENC-NEW	191.91	180.91	162.44	147.10	95.33	138.46	128.64	109.49
ΔR^2_{oos}	0.349	0.335	0.296	0.271	0.704	1.029	0.922	0.661
	Panel A2: $\tilde{G}_t$+$\widetilde{PC}_{1-5,t}$ vs. $\widetilde{PC}_{1-5,t}$							
ENC-REG	4.781	5.118	4.823	4.526	3.654	4.831	5.218	4.829
ENC-NEW	180.94	173.49	151.82	130.10	73.93	134.07	130.97	99.17
ΔR^2_{oos}	0.353	0.340	0.292	0.256	0.809	1.026	0.886	0.543
	Panel B: Economic significance							
	Panel B1: Trading on $\tilde{G}_t$ vs. buy-and-hold							
$\gamma = 3$	0.343	1.267	2.702	4.478	0.308	2.293	4.083	8.745
	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩
$\gamma = 5$	0.565	2.481	5.289	8.622	0.340	4.053	7.858	16.630
	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩
	Panel B2: Trading on $\widetilde{PC}^{o}_{1-3,t} + \tilde{G}_t$ vs. Trading on $\widetilde{PC}^{o}_{1-3,t}$							
$\gamma = 3$	0.432	0.510	0.504	0.449	0.579	1.131	1.133	0.750
	⟨0.008⟩	⟨0.019⟩	⟨0.016⟩	⟨0.012⟩	⟨0.018⟩	⟨0.018⟩	⟨0.009⟩	⟨0.004⟩
$\gamma = 5$	0.292	0.289	0.277	0.239	0.407	0.682	0.685	0.450
	⟨0.022⟩	⟨0.028⟩	⟨0.022⟩	⟨0.019⟩	⟨0.057⟩	⟨0.031⟩	⟨0.009⟩	⟨0.004⟩

Table 3: Finite-Sample Properties of Test Statistics under Spanning Hypotheses I and II

This table presents results based on finite-sample distributions of the statistics that are involved in tests of Spanning Hypotheses I and II (H_0^{S1} and H_0^{S2}). 5,000 bootstrapped samples are generated from spanned term structure models, $SM(\mathcal{L},\mathcal{N})$, specified in Section 5.2.1; the length of each bootstrapped sample is set to be consistent with either the entire data sample (panel A) or the post-1984 data sample (panel B). Results in panels A1 through B2 (panels A3 through B4) are obtained from model $SM(2,3)$ (model $SM(4,5)$) that satisfies H_0^{S1} (H_0^{S2}). Test statistics considered include those computed using the Hansen and Hodrick (1980) GMM covariance estimator (HH) and the Newey and West (1987) HAC covariance estimator (NW) with 18 lags, and the out-of-sample ENC-REG test of Ericsson (1992) and ENC-NEW test of Clark and McCracken (2001). For each set of test statistics, the 95th percentile of the bootstrap distribution is reported as the 5% critical value, and the p-values (in angle brackets) are the frequency of bootstrap replications in which the test statistics are at least as large as the statistic in the data. The "ΔR^2" and "ΔR^2_{oos}" measures denote the incremental R^2 and out-of-sample R^2 of Campbell and Thompson (2008), respectively.

	Panel A: Full sample, 1964–2014				Panel B: Subsample, 1985–2014			
maturity (year)	2	3	4	5	2	5	7	10
	Panel A1: In-sample under H_0^{S1}				Panel B1: In-sample under H_0^{S1}			
HH	4.937	4.896	4.712	4.509	4.080	3.910	3.784	3.594
	⟨0.010⟩	⟨0.003⟩	⟨0.001⟩	⟨0.001⟩	⟨0.003⟩	⟨0.003⟩	⟨0.005⟩	⟨0.005⟩
NW	5.064	5.010	4.839	4.654	3.984	3.867	3.714	3.509
	⟨0.006⟩	⟨0.003⟩	⟨0.001⟩	⟨0.000⟩	⟨0.001⟩	⟨0.000⟩	⟨0.001⟩	⟨0.001⟩
ΔR^2	0.108	0.105	0.099	0.091	0.076	0.080	0.066	0.053
	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩
	Panel A2: Out-of-sample under H_0^{S1}				Panel B2: Out-of-sample under H_0^{S1}			
ENC-REG	4.285	4.195	4.095	3.940	3.421	3.282	3.158	2.996
	⟨0.026⟩	⟨0.018⟩	⟨0.019⟩	⟨0.015⟩	⟨0.045⟩	⟨0.012⟩	⟨0.008⟩	⟨0.004⟩
ENC-NEW	51.03	50.39	47.716	43.622	18.710	17.392	15.596	13.439
	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩
ΔR^2_{oos}	0.167	0.163	0.153	0.139	0.147	0.147	0.122	0.095
	⟨0.000⟩	⟨0.001⟩	⟨0.001⟩	⟨0.001⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩
	Panel A3: In-sample under H_0^{S2}				Panel B3: In-sample under H_0^{S2}			
HH	5.054	4.987	4.909	4.783	4.190	3.947	3.782	3.727
	⟨0.005⟩	⟨0.002⟩	⟨0.002⟩	⟨0.002⟩	⟨0.004⟩	⟨0.003⟩	⟨0.005⟩	⟨0.007⟩
NW	5.202	5.149	5.045	4.962	4.110	3.851	3.745	3.654
	⟨0.003⟩	⟨0.001⟩	⟨0.001⟩	⟨0.000⟩	⟨0.001⟩	⟨0.000⟩	⟨0.002⟩	⟨0.003⟩
ΔR^2	0.117	0.113	0.109	0.103	0.083	0.078	0.070	0.063
	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩
	Panel A4: Out-of-sample under H_0^{S2}				Panel B4: Out-of-sample under H_0^{S2}			
ENC-REG	4.326	4.184	4.120	4.046	3.416	3.325	3.229	3.162
	⟨0.027⟩	⟨0.017⟩	⟨0.019⟩	⟨0.024⟩	⟨0.038⟩	⟨0.009⟩	⟨0.006⟩	⟨0.004⟩
ENC-NEW	56.68	54.53	52.38	49.387	18.516	16.281	15.098	14.050
	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩
ΔR^2_{oos}	0.185	0.177	0.170	0.160	0.161	0.149	0.131	0.119
	⟨0.002⟩	⟨0.003⟩	⟨0.004⟩	⟨0.007⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩

Table 4: Statistical Inference about Unspanned Macro Risks

Panel A reports results from likelihood-ratio tests of the macro-unspanning restrictions (H_0^{US}), given in Eq. (12), that are imposed on an $\mathcal{N}$-factor unconstrained macro-finance term structure model (MTSM). Its underlying state vector is $X_t = (PC_{1-\mathcal{L},t}, \widehat{G}_t)$, where $PC_{1-\mathcal{L},t}$ denotes the vector of the first $\mathcal{L}$ principal components (PCs) of the noise-uncontaminated yield curve and $\widehat{G}_t$ represents the SAGLasso macro factor. Model-based test statistics (column 2) are evaluated against the critical values of a χ^2-distribution with degrees of freedom equal to $(k-\mathcal{N})(\mathcal{N}+1)-1$, where k is the number of bonds involved. Model-free test statistics (column 3) are evaluated based on the $\chi^2(k)$-distribution. The p-values appear in angle brackets immediately beneath. Panel B considers the projection of the SAGLasso macro factor ($\widehat{G}_t$) onto the first $\mathcal{N}$ PCs of the yield curve ($PC^o_{1-\mathcal{N},t}$). Column 5 shows regression R^2s along with two sets of 95% confidence intervals based on 5,000 artificial samples simulated from model $CSM(\mathcal{L},\mathcal{N})$ as specified in Section IA.G.1 (which denotes the $\mathcal{N}$-factor constrained MTSM with a spanned $\widehat{G}_t$ and whose state vector $X_t = (PC_{1-\mathcal{L},t}, \widehat{G}_t)$). The confidence intervals in brackets beneath are obtained under either the assumption that there is no macro measurement error ($\eta_f = 0$) or that there is macro measurement error ($\eta_f \neq 0$), as indicated in column 4 where η_f denotes macro measurement error ("Macro M.E."). Column 6 reports the first-order serial correlation of residuals.

(1)	(2)	(3)	(4)	(5)		(6)
	Panel A: Tests of unspanning restrictions		Panel B: Regressions of $\widehat{G}_t$ on $PC^o_{1-\mathcal{N},t}$			
$\mathcal{N}$	Model-based	Model-free	Macro M.E.	R^2		AR(1) of residuals
4	28.69	10.05		0.145		
	⟨0.122⟩	⟨0.074⟩	No ($\eta_f = 0$)	[0.593	0.847]	
			Yes ($\eta_f \neq 0$)	[0.587	0.769]	0.667
5	24.29	8.23		0.145		
	⟨0.185⟩	⟨0.083⟩	No ($\eta_f = 0$)	[0.506	0.833]	
			Yes ($\eta_f \neq 0$)	[0.459	0.784]	0.667
6	17.55	6.17		0.146		
	⟨0.287⟩	⟨0.104⟩	No ($\eta_f = 0$)	[0.263	0.651]	
			Yes ($\eta_f \neq 0$)	[0.239	0.630]	0.666

Figure 1: Word Cloud from Selected Macroeconomic Series

This figure reports the list of words constituting the names of macroeconomic series that are selected from the SAGLasso algorithm. Font size of a word is proportional to the frequency with which the word appears in selected macroeconomic variables and their lags.

Figure 2: The SAGLasso Factor and Excess Returns on the Five-Year Bond

This figure presents time variation in the normalized SAGLasso factor as well as excess returns on the five-year bond over the sample period from January 1964 to December 2013. Shaded bars denote months designated as recessions by the National Bureau of Economic Research.

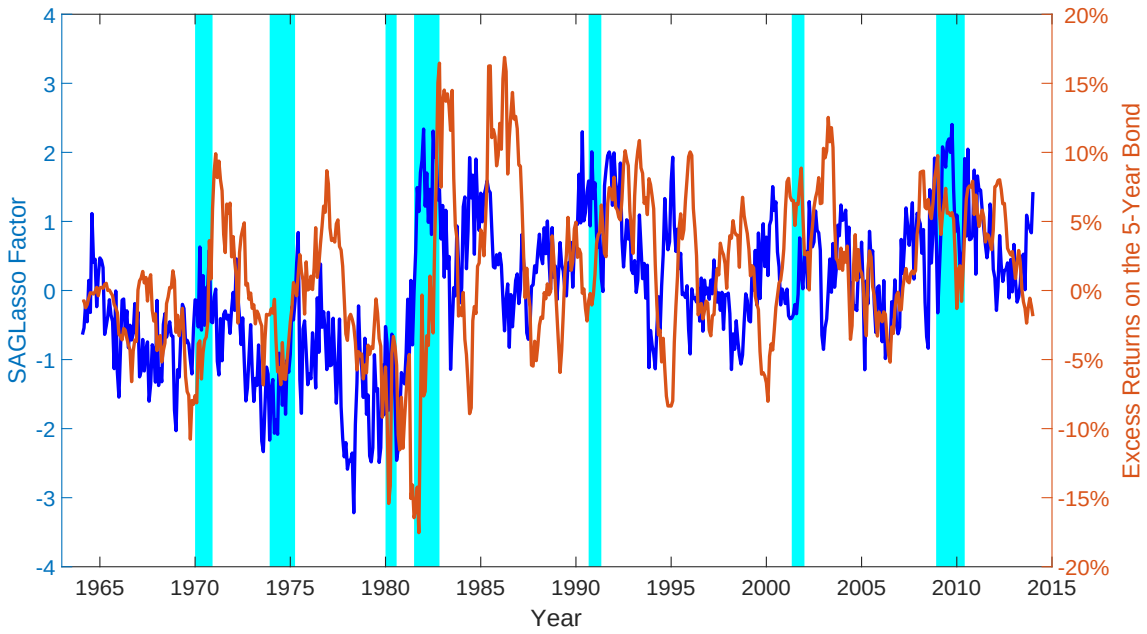
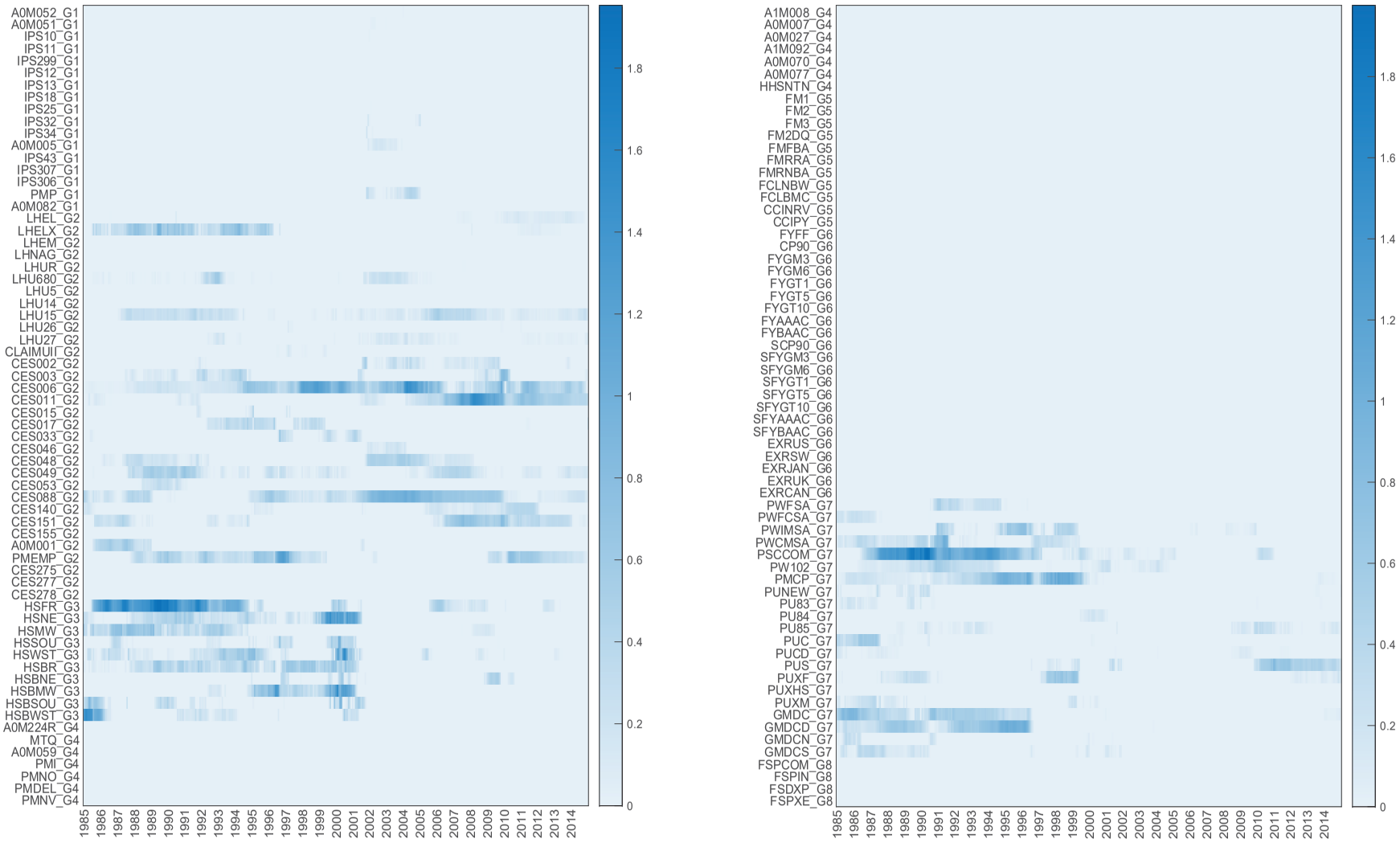


Figure 3: Time Variation of Macro Variable Importance

This figure presents the norm of coefficients associated with the 131 macroeconomic series and their lagged values in the rolling-window bond return prediction. The 131 series are divided into eight groups: G1) output (17 series); G2) labor market (32 series); G3) housing sector (10 series); G4) orders and inventories (14 series); G5) money and credit (11 series); G6) bond and FX—interest rates or financial (22 series); G7) prices or price indices (21 series); and G8) stock market (4 series). In each month since January 1984, the macroeconomic panel data over the past 20 years is input into the SAGLasso algorithm to forecast one-year-ahead excess bond returns, and each macroeconomic series could have at most 7 non-zero coefficients (on their contemporaneous and lagged values). X-axis corresponds to the observation date of excess bond returns, and color gradients within each column indicate the most impactful (dark blue) to least important (white) variables.



Internet Appendix to "Machine-Learning-Based Return Predictors and the Spanning Controversy in Macro-Finance"

Internet Appendix

Table of Contents

List of Tables

List of Figures

IA.A Inferring Higher-Order Yield Principal Components

This section examines the relation between PCs of observed yields and those of "true" yields to provide justification for the use of filtered PCs in tests of Spanning Hypothesis II (H_0^{S2}).

It is known that due to the negligible role of higher-order PCs in the cross section, it is difficult to disentangle them from noise in yields. Panel A of Table IA.A1 illustrates the (limited) effectiveness of direct Principal Component Analysis (PCA) in recovering information in true yields. Column 2 reports population correlations between true yield factors ($PC_{1-5,t}$) and PCs of the observed yield curve ($PC^o_{1-5,t}$), where the correlations are all computed from Monte Carlo simulations based on an estimated five-factor yields-only (Gaussian) term structure model (YTSM). Note that while $\text{corr}(PC^o_{i,t}, PC_{i,t})\,\forall i = 1, 2, 3$ is very high (> 0.96), it is 0.72 for $i = 4$ and 0.21 for $i = 5$, suggesting the inability of PCA to accurately infer $PC_{4,t}$ and $PC_{5,t}$ from yield data. This finding is not surprising given the magnitude of yield loadings on the higher-order PCs: untabulated results indicate that a one-standard-deviation shock to $PC_{4,t}$ or $PC_{5,t}$ does not change any yield by more than seven basis points (bps). On the other hand, the estimated standard deviation of the measurement error in yields is about six bps, which is enough to obscure the cross-sectional effects of $PC_{4,t}$ and $PC_{5,t}$.

In the real data sample, true yield factors are not observable. Duffee (2011) shows that filtering techniques, while no substitute for direct observation, are helpful in retrieving information in those higher-order factors. We find from an unreported simulation analysis that model-implied correlations between true and filtered factors are higher than 0.85 for both the fourth and fifth factors. As a result, we use filtered PCs in our empirical tests of H_0^{S2}.

Will it make a difference at all if we ignore the "hidden" nature of higher-order factors? Column 3 in panel A of Table IA.A1 shows that for these factors, the filtered and PCA-based estimates are significantly different in our 1964–2014 sample.[IA.1] More importantly, replacing the former with the latter leads to underestimation of the predictive power possessed by the historical yield dynamics.

Panel B of Table IA.A1 presents results from regressions of excess bond returns on $PC^o_{1-5,t}$ for two- through five-year maturities. Comparing the panel with Table IA.B5 (columns 10–13)

[IA.1] Joslin, Singleton, and Zhu (2011; Section 6) document very similar results over the sample period 1990–2007: the model-implied filtered $PC_{1-3,t}$ are nearly identical to $PC^o_{1-3,t}$ regardless of the model dimension, but $PC^o_{4-5,t}$ do not closely correspond to their model-implied counterparts. Especially, the authors notice that filtered high-order factors appear to be a smoothed version of $PC^o_{4-5,t}$.

reveals that replacing (filtered) $PC_{4-5,t}$ by $PC^o_{4-5,t}$ results in lower R^2 values, regardless of the bond maturity, where for convenience those four R^2 values from Table 2 are shown in the row labeled "R^2 (Table 2)" in panel B. Note that the decline in R^2 ranges from 0.014 for the four-year bond to 0.018 for the three-year bond (the second last row of panel B), and the percentage decrease in R^2 ranges from 5.49% for the four-year bond to 7.73% for the three-year bond (the last row). Given that the first three factors alone can have an R^2 of 16–19% (columns 2–5 of Table IA.B5), the aforementioned amount of information loss in the fourth and fifth factors is far from trivial. As such, using $PC^o_{1-5,t}$ in regression tests of H_0^{S2} would make the hypothesis overrejected.

IA.B More on Properties of the SAGLasso Macro Factor

This appendix further examines the properties of the SAGLasso macro factor ($\widehat{G}$). Section IA.B.1 investigates the predictive power of the three group factors which constitute the SAGLasso factor. Section IA.B.2 examines whether the three group factors are spanned or not. Section IA.B.3 compares the predictive power of the $\widehat{G}$ and Ludvigson and Ng (2009; LN09 hereafter) macro factors. Section IA.B.4 investigates the potential impact of data revision and publication lags on return predictability. Section IA.B.5 presents in-sample tests of the spanning hypotheses, H_0^{S1} and H_0^{S2}. Lastly, Section IA.B.6 implements the SAGLasso algorithm using 131 macro variables along with different numbers of their lags.

IA.B.1 Predictive Power of the Three Group Factors

The SAGLasso macro factor $\widehat{G}_t$ consists of three group factors: the employment ($\widehat{g}_{1t}$), housing ($\widehat{g}_{2t}$), and inflation ($\widehat{g}_{3t}$) factors. To better understand the information content of factor $\widehat{G}_t$, we examine properties of these three group factors in this subsection. Let $\{\widehat{g}_{it}\}_{1\leq i\leq 3}$ denote $\{\widehat{g}_{it}, 1 \leq i \leq 3\}$.

IA.B.1.1 Sample Period 1964–2014

Table IA.B1 reports the Pearson correlation matrix of $\widehat{G}$, $\{\hat{g}_{it}\}_{1\leq i\leq 3}$, and five yield curve factors. The five yield curve factors include the first three principal components (PCs) of *observed* bond yields, $\{PC^o_{i,t}, i = 1, 2, 3\}$, and the filtered higher-order PCs of noise-uncontaminated yields, $PC_{4,t}$ and $PC_{5,t}$. As expected, $\widehat{g}_{1t}$, $\widehat{g}_{2t}$, and $\widehat{g}_{3t}$ all have low correlations with the yield curve factors. In

particular, the novel housing factor $\widehat{g}_{2t}$ has a correlation of -0.167 with $PC^o_{1,t}$, -0.073 with $PC^o_{2,t}$, and 0.222 with $PC^o_{3,t}$. As a result, $\widehat{G}$ is weakly correlated with $PC^o_{1-3,t}$ and hardly correlated with $PC_{4,t}$ and $PC_{5,t}$. Recall that by construction the $\widehat{G}$ factor and its three component factors control for the Treasury and FX variables (group 5) out of the 131 macro series. The results shown in the table verify that $\widehat{G}$ and $\{\hat{g}_{it}\}_{1\leq i\leq 3}$ are all weakly correlated with the yield curve. Nonetheless, as shown below these group factors have strong predictive power as a result of the SAGLasso procedure used for model selection.

We now examine the predictive power of $\widehat{g}_{1t}$, $\widehat{g}_{2t}$, and $\widehat{g}_{3t}$, both individually and jointly. Panel A of Table IA.B2 presents results from predictive regressions of excess bond returns on normalized $\widehat{g}_{1t}$, $\widehat{g}_{2t}$, and $\widehat{g}_{3t}$, for 2-, 3-, 4-, and 5-year bonds in the full sample period. Panel A1 reports coefficient estimates, t-statistics, and R-squared of univariate regressions on each of the three group factors. Note that these factors all exhibit significant unconditional predictive power, with an R^2 of 21–22% for $\widehat{g}_{1t}$, about 14–15% for $\widehat{g}_{2t}$, and 17–18% for $\widehat{g}_{3t}$. Results from multivariate regressions, reported in panel A2, show that the three group factors are still all significant and together have an (adjusted) R^2 ranging from about 40% for the 2-year bond to 43% for the 3-year bond.

As shown in Joslin, Priebsch, and Singleton (2014; JPS hereafter), the impact of macro risk factors on bond risk premia depends on horizons. Panel A of Table IA.B2 illustrates the relative importance of the three group factors across bond maturity. The univariate regression results indicate that the regression coefficient on $\widehat{g}_{1t}$ is the largest, followed by the one on $\widehat{g}_{3t}$, and the coefficient on $\widehat{g}_{2t}$ is the lowest, regardless of the bond maturity. The univariate regression R^2 values exhibit the same pattern. In the multivarite regressions, the regression coefficients on the three group factors maintain the same ranking, regardless of the bond maturity. These results indicate that relatively speaking, among the three group factors, the employment factor ($\widehat{g}_{1t}$) is the most important, followed by the inflation factor ($\widehat{g}_{3t}$), and then by the housing factor ($\widehat{g}_{2t}$). Note, however, that these group factors are trained on the aggregate bond market returns rather than returns on bonds with a specific maturity. Bianchi, Büchner, and Tamoni (2021) consider more categories of macro variables and find that variables related to the stock and labor market (the output & income and orders & inventories) are more important for the short-end (long-end) of the yield curve.

IA.B.1.2 Sample Period 1952–2014

The full sample used in this study is 1964–2014. However, it is known that the relationship between interest rate and real activity changed significantly around 1964. This raises one concern about the robustness of our evidence for the predictive power of $\widehat{g}_{it}, i = 1, 2, 3$ and $\widehat{G}_t$ based on the 1964–2014 sample: If we extend the sample to several years earlier, that may significantly change the results. To address this concern, we reexamine the predictive power of these macro factors using the sample extended to 1952, the year from which the data coverage of the original Fama-Bliss yields starts.[IA.2] However, some macro series, especially those related to housing, are not available going back that far; thus, we reconstruct the employment factor only in this robustness check, and denote the factor constructed in-sample by $\widehat{g}^*_{1t}$ and its out-of-sample version by $\widetilde{g}^*_{1t}$. Recall from Section 3 that the "labor" group includes 32 series. As two of these series are no longer available when the sample extends back to 1952, $\widehat{g}^*_{1t}$ or $\widetilde{g}^*_{1t}$ is constructed using the remaining 30 macro series.

Results from in-sample regressions, reported in panel B1 of Table IA.B2, indicate that the predictive power of the employment factor is robust to the extended sample. Comparing panels A1 and B1, we see that the predictive power of $\widehat{g}^*_{1t}$ is slightly weaker than that of $\widehat{g}_{1t}$ in terms of the magnitude of regression coefficients or R^2 value but the coefficient on $\widehat{g}^*_{1t}$ has greater t-value than that on $\widehat{g}_{1t}$, regardless of the bond maturity.

In the out-of-sample tests, the training period is 20 years, which is close to the 21-year period length adopted in our full-sample (1964–2014) analysis. In other words, the employment factor is reconstructed every month after December 1971 using Adaptive Lasso within a given group, and the predictive regression is also reestimated recursively. As before, we consider the following three out-of-sample statistics: the "ENC-REG" (Ericsson 1992), the "ENC-NEW" (Clark and McCracken 2001), and the out-of-sample R-squared "R^2_{oos}" (Campbell and Thompson 2008) statistics. The results shown in panel B2 of Table IA.B2 indicate that $\widetilde{g}^*_{1t}$ has significant out-of-sample predictive power for every bond considered. Additionally, R^2_{oos} increases in the bond maturity, ranging from 0.155 for the 2-year bond to 0.169 for the 5-year bond.

Overall, the above results provide evidence that the predictive power of the employment factor (one main component of the SAGLasso single factor $\widehat{G}_t$) is robust to the longer sample 1952–2014.

[IA.2] The supplement to Cochrane and Piazzesi (2005), available at `http://www.stanford.edu/~piazzesi/cp.zip`, suggests that Fama-Bliss yield data prior to 1964 is unreliable.

IA.B.2 Spanning Properties of the Group Factors

Having examined the predictive power of the three group factors, we explore, to what extent, each of the three factors is spanned or unspanned in this subsection.

Recall from Table IA.B1 that $\widehat{g}_{1t}$, $\widehat{g}_{2t}$, and $\widehat{g}_{3t}$ all have low correlations with the yield curve factors. In an untabulate analysis, we find that the three group factors are not highly correlated with *GRO* (growth) and *INF* (inflation) either, two standard single macro variables used in the literature. Unsurprisingly, the largest correlation (0.497) occurs between the two inflation factors, $\widehat{g}_{3t}$ and INF_t. The correlations between INF_t and the other two group factors are 0.237 for $\widehat{g}_{1t}$ and 0.144 for $\widehat{g}_{2t}$. The growth variable *GRO* has a correlation of -0.013, 0.167, and -0.015 with $\widehat{g}_{1t}$, $\widehat{g}_{2t}$, and $\widehat{g}_{3t}$, respectively. These findings suggest that the three group factors are viable candidates for unspanned macro variables.

We examine whether the three group factors are spanned by the yield curve, following Section 5.4 that conducts a similar analysis for $\widehat{G}_t$. That is, for each of the three group factors, we first regress the factor on the first $\mathcal{R}$ PCs of the yield curve ($PC^o_{1-\mathcal{R},t}$), where $\mathcal{R} = 3$ or 6; we then evaluate the regression R^2 against its distribution implied from a constrained and spanned model; we also estimate the first-order correlation of residuals from the regression to see if the residuals are serially uncorrelated. The model used here to generate distributions of R^2 is model $CSM(3,6)_{group}$, whose state vector includes three yield curve factors (the first three PCs) and three macro factors, $\widehat{g}_{1t}$, $\widehat{g}_{2t}$, and $\widehat{g}_{3t}$. The estimation of the model is done under the assumption that the three macro factors are measured either with or without errors.

Table IA.B3 reports the regression results for each of the group factors with $\mathcal{R} = 3$ (panel A) or 6 (panel B). Column 2 indicates whether the three macro variables are assumed to be measured with errors in the estimation of model. Columns 3 (panel A) and 5 (panel B) show the regression R^2s, and in brackets beneath are reported 95% confidence intervals based on 5,000 artificial samples simulated from model $CSM(3,6)_{group}$. Columns 4 (panel A) and 6 (panel B) report the first-order serial correlation of regression residuals. Clearly, the regression R^2 is outside of the 95% confidence intervals for each of the group factors in either panel. Moreover, even the smallest estimated first-order serial correlation is around 90%, suggesting that much of the persistent component is mistakenly treated as white-noise shocks. All of the evidence indicates that the three SAGLasso

group macro factors $\{\widehat{g}_{it}, 1 \leq i \leq 3\}$ are not spanned by the yield curve factors.

IA.B.3 Comparison with the Ludvigson and Ng (2009) Factor

The LN09 single factor, constructed through dynamic factor analysis and BIC-based stepwise predictive regression, is $\overrightarrow{F6}_t = (\widehat{F}_{1t}, \widehat{F}^3_{1t}, \widehat{F}_{2t}, \widehat{F}_{3t}, \widehat{F}_{4t}, \widehat{F}_{8t})$, the particular polynomial function of LN09's eight dynamic factors that minimizes the BIC over the sample period 1964–2003. However, using our panel of 131 "real-time" macro series over 1964–2014, we find that the selected subset includes $\overrightarrow{\mathcal{F}7}_t = (\widehat{F}_{1t}, \widehat{F}^3_{1t}, \widehat{F}_{2t}, \widehat{F}_{5t}, \widehat{F}^2_{5t}, \widehat{F}_{8t}, \widehat{F}^2_{8t})$,[IA.3] whose R^2 value is 0.256, higher than 0.214 of $\overrightarrow{F6}_t$'s. Hence, we let $\widehat{LN}^m_t$ (the modified LN factor) be $\overrightarrow{\mathcal{F}7}_t$ in our empirical analysis.

Although both $\widehat{G}$ and $\widehat{LN}^m$ are extracted from the same set of 131 macro series, they differ in several aspects. First, whereas $\widehat{LN}^m$ includes all 131 series and squares and cubes of these macro variables, $\widehat{G}$ is a *linear* combination of 19 series and some of their lagged variables, and consists of three easy-to-interpret macro group factors. Second, in terms of economic interpretation, $\widehat{G}$ includes a housing factor that contributes little to $\widehat{LN}^m$, whose important components are the "real activity" (highly correlated with measures of employment and production such as IP growth), "inflation," and "stock market" factors. Also, $\widehat{G}$ includes no variables from the "bond and FX" group and thus is much less correlated with the yield curve than $\widehat{LN}^m$ is. Lastly, by construction $\widehat{G}$ takes into account the dynamic response of bond risk premia to macroeconomic innovations. In contrast, information on term premia does not enter $\widehat{LN}^m$ until the last step of its construction.

Panel A of Table IA.B4 reports the prediction results based on $\widehat{LN}^m_t$ for the full sample. Results from the in-sample analysis reported in panel A1 show that $\widehat{LN}^m_t$ is significant, regardless of the bond maturity, and that the R^2 increases in the bond maturity, ranging from 0.168 for the 2-year bond to 0.250 for the 5-year bond. Recall from panel A1 of Table 1 that the R^2 from regressions on $\widehat{G}_t$ ranges from 0.352 for the 2-year bond to 0.392 for the 5-year bond. The difference between this R^2 and that of $\widehat{LN}^m_t$ is 0.18, 0.16, 0.15, and 0.14 for the 2-, 3-, 4-, and 5-year bonds, respectively. These results indicate that $\widehat{G}_t$ has a greater predictive power than $\widehat{LN}^m_t$ for excess bond returns.[IA.4] Results from the out-of-sample analysis also support this conclusion, as can be seen from evidence shown in panel A2 of Table IA.B4 for $\widehat{LN}^m_t$ and that in panel A of Table 2 for $\widehat{G}_t$. To summarize,

[IA.3] The variable $\widehat{F}^2_{5t}$ is also selected by Ludvigson and Ng (2011), who consider the sample period 1964–2008.
[IA.4] This finding is robust in the post-1984 sample period (untabulated).

even though $\widehat{G}_t$ is linear and much more parsimonious than $\widehat{LN}_t^m$, the former predictor shows stronger predictive ability than the latter in both in-sample and out-of-sample analyses.

IA.B.4 Data Revisions, Publication Lags, and Return Predictability

The SAGLasso macro factor $\widehat{G}$ (as well as $\widehat{LN}^m$ considered before) is constructed based on the set of 131 macro series compiled in this study that adjust for both data revisions and publication lags. This subsection examines the impact of these two adjustments on bond return predictability.

To this end, we construct two new macro factors using the same SAGLasso procedure as described before in Section 4 but with different macro data. The first factor, denoted $\widehat{G}_t^{rev}$, is constructed based on the set of the same 131 macro series that, however, adjust for publication lags only (and not data revisions). The other new macro factor, denoted $\widehat{G}_t^{rev,lag}$, is constructed based on the set of the 131 macro series that does not adjust for either data revisions or publication lags—namely, the original set of macro series used in LN09 less the one series no longer available.

Panel B of Table IA.B4 reports the results from predictive regressions of excess bond returns on $\widehat{G}_t^{rev}$ from both in-sample (panel B1) and out-of-sample (panel B2) analyses. Comparing panel B1 with panel A1 of Table 1 reveals that both the regression coefficient on $\widehat{G}_t^{rev}$ and its in-sample R^2 are slightly larger than those for $\widehat{G}_t$ except for the 2-year bond. Similarly, the out-of-sample R^2_{oos} of $\widehat{G}_t^{rev}$ (panel B2 of Table IA.B4) is slightly higher than that of $\widehat{G}_t$ (panel A of Table 2), regardless of the bond maturity. These findings indicate that the predictive ability of $\widehat{G}_t$ is slightly inferior to that of $\widehat{G}_t^{rev}$. In other words, data revisions inflate the predictability only slightly in our sample.

Conversely, results reported in panel C of Table IA.B4 show that return predictability is substantially exaggerated if publication lags are not adjusted. For instance, the in-sample R^2 of $\widehat{G}_t^{rev,lag}$ is 0.414, 0.441, 0.453, and 0.464 for the 2- through 5-year bonds, respectively (panel C1) and is much higher than that of $\widehat{G}_t$ (panel A1 of Table 1). The increase in the R^2 ranges from 6.2% for the 2-year bond to 7.2% for the 5-year bond. That is, the inflated predictability is especially notable in the in-sample regressions. The out-of-sample evidence shown in panel C2 of Table IA.B4 (based on $\widehat{G}_t^{rev,lag}$) and panel A of Table 2 (based on $\widehat{G}_t$) also indicates that ignoring publication lags inflates the predictability, albeit to a lesser degree.

To summarize, we find that publication lags pose much greater "danger" than data revisions in forecasting future bond returns based on macro variables, at least in our sample. This problem can

be mitigated straightforwardly, however, since in practice it is easier to make an adjustment for publication lags than to figure out preliminary macro data releases and adjust for data revisions.

Note that the main finding of this subsection is consistent with Ghysels, Horan, and Moench (2018), who document that using revised macro series inflates the predictive power of macro variables. However, while they focus on a particular macro variable—"total non-farm payroll employment" (#33 on our list of 131 series)—and find that both data revisions and publication lags are highly important, we examine the impact of these two elements on a large panel of macro time series and find that the predictive power of the SAGLasso (aggregate) macro variable is robust to the use of vintage data. Namely, the importance of revision/delay biases depends on specific macro series, especially given that variable #33 itself is not included in $\widehat{G}$ (see Table A.1 in the paper). This implication is consistent with Ghysels et al. (2018) too. In a robustness analysis, they consider the Chicago Fed National Activity Index (an unsmoothed version of macro variable *GRO*) and find that the combined effect of publication lags and data revisions on these two aggregate macro variables is small. Also, Barillas (2012) finds that the bond return predictability is robust to the use of real time series for 16 macro variables (7 inflation and 9 real growth measures) considered in his study.

IA.B.5 In-Sample Spanning Tests

This subsection tests the spanning hypotheses, H_0^{S1} and H_0^{S2}, by examining the incremental predictive power of $\widehat{G}$ over the yield curve. As before, we focus mainly on the test statistics based on the HH or NW standard errors in the discussion of test results that follows.

Table IA.B5 presents the results based on the full sample. Results from regressions on $PC^o_{1-3,t}$, reported in columns 2–5, indicate that only $PC^o_{2,t}$ is significant and that the R^2 ranges from 0.156 for the 3-year bond to 0.194 for the 5-year bond. Results from each of the above regressions augmented with $\widehat{G}_t$, reported in columns 6–9, show that $\widehat{G}$ is significant regardless of the bond maturity. The incremental R-squared due to $\widehat{G}$, ΔR^2, ranges from 0.243 for the 5-year bond to 0.262 for the 3-year bond. These results provide strong evidence against H_0^{S1}.

Results from regressions on $PC_{1-5,t}$, shown in columns 10 through 13, indicate that in addition to $PC_{2,t}$, the higher-order $PC_{4,t}$ and $PC_{5,t}$ are also significant for most bonds.[IA.5] The R^2 ranges

[IA.5] Internet Appendix IA.A presents empirical evidence that the PCA of the observed yields is unable to effectively

from 0.221 for the 2-year bond to 0.255 for the 4-year bond. Augmenting these regressions with $\widehat{G}_t$ yields a ΔR^2 ranging from 0.232 for the 5-year bond (column 17) to 0.253 for the 3-year bond (column 15). Importantly, $\widehat{G}_t$ is significantly different from zero regardless of the bond maturity and standard errors used, indicating a rejection of H_0^{S2}. In addition, $PC_{2,t}$ and $PC_{4,t}$ become less significant (and insignificant for the 2- and 3-year bonds) in the presence of $\widehat{G}_t$.

The results for the post-1984 sample, reported in Table IA.B6, are qualitatively the same as those for the full sample. Particularly, $\widehat{G}_t$ is significant, regardless of the bond maturity and standard errors used, conditional on either $PC^o_{1-3,t}$ (columns 6–9) or $PC_{1-5,t}$ (columns 14–17); namely, H_0^{S1} and H_0^{S2} are strongly rejected by the post-1984 sample too. Compared with its counterparts for the full sample (Table IA.B5), ΔR^2 due to $\widehat{G}_t$ is actually higher except for the 2-year bond. For instance, ΔR^2 from regression tests of H_0^{S1} for the 5-year bond is 0.297 for the post-1984 sample (column 7) and 0.243 for the full sample (column 9 in Table IA.B5). Regarding the impact of PCs in the presence of $\widehat{G}_t$, $PC^o_{1,t}$ ($PC^o_{2,t}$) remains significant for the 2- and 5-year bonds (10-year bond) in the tests of H_0^{S1}. For regression tests of H_0^{S2} (columns 14–17), $PC_{1,t}$ is significant regardless of the bond maturity, $PC_{2,t}$ is significant for the 7- and 10-year bonds, and $PC_{5,t}$ for the 2-year bond only.

An earlier version of the paper also considers test statistics based on Hodrick 1B standard errors. We find that $\widehat{G}$ remains significant regardless of the bond maturity, whereas some of the PCs become insignificant. For instance, $PC^o_{2,t}$ remains significant only for the 4- and 5-year bonds and is subsumed by $\widehat{G}_t$ regardless of the bond maturity.

In summary, when factor $\widehat{G}$ is used as the macro-based return predictor, our in-sample test results show that this new macro variable has predictive power above and beyond the contemporaneous yield curve or yield dynamics, and thereby reject both Spanning Hypotheses I and II.

IA.B.6 Tests Using Macro Variables with Different Lags

So far the SAGLasso algorithm has been implemented using 131 macro variables along with six of their lags. In this subsection we address the following two questions: (1) Are lags of macro variables are essential to maintain the predictive performance as documented in Section 4, given

disentangle higher-order PCs from noise in yields. Filtered higher-order PCs ($PC_{4-5,t}$) contain more information about bond risk premia than higher-order observed PCs ($PC^o_{4-5,t}$).

that 21 constituent variables (out of 30) of G are lagged? (2) If so, what is the optimal number of lags to be included in our supervised learning?

These are nontrivial questions as a panel of macro data with no lags or a small number of lags has a denser structure and might deliver better out-of-sample performance given the limited length of the training period. To see this, recall that tuning parameters are selected using cross-validations in the SAGLasso algorithm (see Appendix C). Therefore, as we include more and more lags, the estimation process is inevitably subject to more "noise", which could overweigh benefits of incorporating more historical information in the construction of the SAGLasso factor.

In what follows, we repeat the analysis in Section 4.1 using 131 macro variables along with N_L of their lags, where $N_L = 0, 3, 9, 12$. To be more specific, for each value of N_L, we first reconstruct the SAGLasso factor following the procedures described in Appendix C and then examine the predictive power of the reconstructed SAGLasso factor.

Figure IA.B1 depicts the unconditional predictive power of the SAGLasso factor constructed using the macro data with $N_L = 0, 3, 6, 9, 12$. For brevity, we report the results for 2-year and 5-year bonds only. Panel A shows that including lags clearly enhances the in-sample predictive power of the SAGLasso macro factor.[IA.6] However, using more lags does not necessarily raise the R^2 value: it is the highest with $N_L = 3$ for the 2-year bond and with $N_L = 6$ for the 5-year bond.

As discussed above, including more than 6 lags may induce nontrivial sampling variability of the SAGLasso estimates that is sufficiently large to offset the gains from using more data. This conjecture is confirmed by the results for the out-of-sample R^2 shown in Panel B. Since the SAGLasso factor is estimated recursively (with a rolling 20-year window) in the out-of-sample analysis, we face greater uncertainty compared to the in-sample estimation. As a result, we find that the SAGLasso factor with $N_L = 9$ or 12 hardly outperforms the SAGLasso factor with $N_L = 0$ (no lag) in terms of the out-of-sample R^2.

Overall, the results shown in Figure IA.B1 suggest that the SAGLasso factor constructed using the 131 macro variables along with 3 or 6 of their lags has the best performance in both the in-sample and out-of-sample predictions. This finding reflects a trade-off between including more information

[IA.6] Note that including lags into the SAGLasso algorithm does not simply lead to an expansion in the set of selected macro variables. Instead, the coefficients of some previously selected (contemporaneous) variables are shrunk to zero, "crowded out" by more powerful lagged variables. For example, 29 macro variables are selected with $N_L = 0$, but only 9 of them have nonzero coefficients with $N_L = 3$.

in the supervised learning and imposing a denser data structure to enhance the estimation stability. While the baseline SAGLasso factor (with $N_L = 6$) seems to capture more information on long-term bond premiums, the alternative SAGLasso factor with $N_L = 3$ outperforms for short-term bonds.

Next, we examine whether or not the choice of lag length affects our inferences with respect to Spanning Hypotheses I and II. We test these two hypotheses using the above SAGLasso factor with different values of N_L and report the test results in panels A and B of Table IA.B7, respectively. As before, the test statistics used include the Hansen-Hodrick one, the Newey-West statistic, and ΔR^2 (the incremental in-sample R^2) for the in-sample tests, as well as the ENC-REG statistic, the ENC-NEW statistic, and ΔR^2_{oos} (the incremental out-of-sample R^2). Note that both of the spanning hypotheses are overwhelmingly rejected in both the in-sample and out-of-sample tests, regardless of the value of N_L considered. In particular, the two hypotheses are strongly rejected when no lags ($N_L = 0$) are used in the construction of the SAGLasso factor.

Finally, we perform the finite-sample analysis based on the SAGLasso factor with $N_L = 3$. Untabulated results show that the finite-sample critical values of the aforementioned six statistics are fairly close to their counterparts as reported in Table 4. It follows that this newly formed macro factor still results in a rejection of the two spanning hypotheses. Therefore, an alteration to the lag length does not change the conclusion on the finite-sample tests.

To summarize, our test results indicate that the choice of lag length hardly affects the our inferences with respect to the two spanning hypotheses.

IA.C Estimation and Selection of MTSMs

It is mentioned in Section 5.2.2 that in our estimation of MTSMs we use the canonical form of Gaussian MTSMs developed by Joslin, Le, and Singleton (2013; hereinafter JLS). This section reviews the JLS canonical form first. We then discuss restrictions on risk premium parameters.

IA.C.1 The Joslin, Le, and Singleton (2013) Canonical Form

We follow the JPS framework for MTSMs in Section 5.1. However, for the purpose of estimation, it is convenient to use a slightly different parameterization that is consistent with the JLS canonical form (following JPS and Duffee 2013a). Building on the Joslin, Singleton, and Zhu (2011) canonical

form for YTSMs, the JLS canonical form defines the most general admissible Gaussian MTSM for a given dimension of the state vector.

Denote the state vector satisfying the JLS canonical form is denoted by X_t^*. Its $\mathbb{Q}$-measure dynamics and the resulting bond pricing formula are

$$r_t = r_\infty^{\mathbb{Q}} + \imath \cdot X_t^*, \tag{IA.C1}$$

$$X_t^* = \Phi_x^{*\mathbb{Q}} X_{t-1}^* + \Sigma_x^* \epsilon_t^{\mathbb{Q}}, \tag{IA.C2}$$

$$y_t^{(m)} = A_m^*(\Theta_Y^{\mathbb{Q}}) + B_m^*(\Theta_Y^{\mathbb{Q}})' X_t^* \tag{IA.C3}$$

where $r_\infty^{\mathbb{Q}}$ denotes the long-run mean of the short rate under $\mathbb{Q}$,[IA.7] $\imath$ is a vector of ones, $\Phi_x^{*\mathbb{Q}} - I$ has the real Jordan form determined by the eigenvalue vector $\gamma^{\mathbb{Q}}$, and Σ_x^* is lower triangular. Under this representation, $\Theta_Y^{\mathbb{Q}} \equiv \{\gamma^{\mathbb{Q}}, r_\infty^{\mathbb{Q}}, \Sigma_x^*\}$ governs X_t^*'s $\mathbb{Q}$-dynamics and thus fully determines bond pricing. Coefficients A_m^* and B_m^* are given by

$$\begin{aligned} B_m^* &= \frac{1}{m}\left(I - \Phi_x^{*\mathbb{Q}\prime}\right)^{-1}\left(I - (\Phi_x^{*\mathbb{Q}\prime})^m\right)\imath, \\ A_m^* &= r_\infty^{\mathbb{Q}} - \frac{1}{2m}\sum_{i=1}^{m-1} B_i^{*\prime}\Sigma_x^*\Sigma_x^{*\prime}B_i^*. \end{aligned}$$

While the state vector X_t^* defines the minimum number of parameters shaping the risk-neutral distribution of bond yields, it keeps silent about the role of macro factors F_t in bond pricing. Unless the macro-unspanning restrictions, as specified in Eq. (12), are imposed, F_t are included in MTSMs as pricing factors, i.e., there is a linear mapping between F_t and X_t^* as follows:

$$F_t = A_f + B_f X_t^*.$$

For ease of notation, in the discussion that follows in this this subsection, we drop the subscript/superscript $\mathcal{M}$ from $Y_t^{\mathcal{M}}$ and $\{\mathcal{A}_{\mathcal{M}}^*, \mathcal{B}_{\mathcal{M}}^*\}$, where $\mathcal{M}$ denotes the maturities of zero yields to be considered. Suppose that the yield-curve factors in X_t are defined by a full-rank loading matrix

[IA.7] In the JSZ canonical form there is no constant term in the short-rate equation (IA.C1). Instead, there is a constant term in the transition equation:

$$X_t^* = \mu_x^{*\mathbb{Q}} + \Phi_x^{*\mathbb{Q}} X_{t-1}^* + \Sigma_x^* \epsilon_t^{\mathbb{Q}},$$

where $\mu_x^{*\mathbb{Q}} = (u_\infty^{\mathbb{Q}}, 0_{1\times(\mathcal{N}-1)})'$. However, as long as X_t^* is stationary under the risk-neutral measure and the first element of $\gamma^{\mathbb{Q}}$ is non-repeated, $r_\infty^{\mathbb{Q}}$ and $u_\infty^{\mathbb{Q}}$ are interchangeable in defining the canonical form: $r_\infty^{\mathbb{Q}} = -u_\infty^{\mathbb{Q}}/\gamma_1^{\mathbb{Q}}$.

$W_{\mathcal{L}} \in \mathbb{R}^{\mathcal{L}\times\mathcal{L}}$, i.e., $\mathcal{P}_t = W_{\mathcal{L}} Y_t$. It follows that the latent state vector X_t^* can be rotated to X_t[IA.8]

$$X_t = \Gamma_0 + \Gamma_1 X_t^*,$$

where

$$\Gamma_0 = \begin{bmatrix} W_{\mathcal{L}}\mathcal{A}^* \\ A_f \end{bmatrix}, \Gamma_1 = \begin{bmatrix} W_{\mathcal{L}}\mathcal{B}^* \\ B_f \end{bmatrix}.$$

The resultant bond-pricing coefficients for the rotated state vector X_t are

$$\begin{aligned} B_m &= \Gamma_1^{-1\prime} B_m^*, \\ A_m &= A_m^* - B_m \Gamma_0. \end{aligned}$$

This leads to a closed-form expression for the probability density function of observed yields conditional on X_t, which completes the maximum likelihood estimation.

Note that $\{A_m, B_m\}$, defined in Eq. (8), depend on $\Theta_M^{\mathbb{Q}} \equiv \{\gamma^{\mathbb{Q}}, r_\infty^{\mathbb{Q}}, A_f, B_f, \Sigma_x^*\} \supset \Theta_Y^{\mathbb{Q}}$. As such, adding macro factors to DTSMs allows for greater flexibility in fitting the conditional distribution of bond yields, as evidenced by the $(\mathcal{N}-\mathcal{L})(\mathcal{N}+1)$ additional free parameters in MTSMs.[IA.9]

Even if we ignore the additional flexibility offered by F_t, it is preferable to factorize the conditional likelihood function in terms of $X_t = (\mathcal{P}_t', F_t')$, as opposed to latent factors X_t^*. First, if the yield portfolios as represented by $\mathcal{P}_t$ are assumed to be priced perfectly (JSZ; JPS), the $\mathbb{P}$-measure conditional density of state variables, $l(X_t|X_{t-1}, \mu_x^{\mathbb{P}}, \Phi_x^{\mathbb{P}}, \Sigma_x)$, can be assessed with standard linear projection; JSZ show that the OLS leads to ML estimators of $\{\mu_x^{\mathbb{P}}, \Phi_x^{\mathbb{P}}\}$. Second, even if we allow all yields to be measured with error, an OLS regression of X_t^o $(= X_t + \eta_t)$ provides fairly reasonable starting values in the estimation of $\{\mu_x^{\mathbb{P}}, \Phi_x^{\mathbb{P}}, \Sigma_x\}$.

IA.C.2 Selection of MTSMs

In Sections IA.G.2 and IA.G.3 of the paper, we follow JPS and conduct a large-scale search for the best set of zero restrictions on risk premium parameters in constrained models $CSM(\mathcal{L},\mathcal{N})$ and $CUSM(\mathcal{L},\mathcal{N})$. This section provides details of this analysis.

[IA.8] The invariant transformation from X_t^* to X_t calls for the loading matrix $W_{\mathcal{L}}$. As the number of yield factors $\mathcal{L} \leq 5$ in models considered in Sections 5.2 and IA.G, $W_{\mathcal{L}}$ is estimated based on model $YTSM(5)$ (see Internet Appendix IA.A for details). Unreported results show that the first three rows of W_5 are almost identical to those of W_5^o (as well as the loading matrix implied from model $YTSM(3)$), but there is substantial difference in the remaining rows.

[IA.9] Therefore, model $SM(\mathcal{L},\mathcal{N})$ has $2.5\mathcal{N}^2 + 3.5\mathcal{N} - \mathcal{N}\mathcal{L} - \mathcal{L} + 2$ parameters in total to estimate.

Recall from Section IA.G.1 that $CSM(\mathcal{L},\mathcal{N})$ and $CUSM(\mathcal{L},\mathcal{N})$ denote $\mathcal{N}$-factor constrained, spanned and unspanned MTSMs, respectively, where the underlying state vector $X_t = (PC_{1-\mathcal{L},t}, \widehat{G}_t)$ and $PC_{1-\mathcal{L}} = (PC_1, \ldots, PC_{\mathcal{L}})$ denotes the first $\mathcal{L}$ PCs of bond yields. The one-period risk premium is as specified in Eq. (13):

$$\Sigma\Lambda_t = \lambda_0 + \lambda_1 X_t \;=\; \lambda_0 + \lambda_1 \cdot (PC_{1-\mathcal{L},t},\; \widehat{G}_t)',$$

where risk premium parameters λ_0 and λ_1 are an $\mathcal{N}$-dimensional vector and an $\mathcal{N} \times \mathcal{N}$ matrix, respectively. In the discussion below, we focus on the selection of spanned models $CSM(\mathcal{L},\mathcal{N})$. The selection of unspanned models is done similarly.

Table IA.C1 shows the maximum likelihood estimates of λ_0 and λ_1 in models selected by BIC, under $CSM(3,4)$ (panel A), $CSM(4,5)$ (panel B), and $CSM(5,6)$ (panel C), respectively. Note from the three panels that while the estimates are model dependent, they show three robust properties that hold regardless of the model dimension $\mathcal{N}$.[IA.10] First, both $\lambda_1(1,\mathcal{N})$ and $\lambda_1(2,\mathcal{N})$ are negative and statistically significant, $\forall \mathcal{N} = 4,5,6$. For instance, $\lambda_1(1,4) = -6.11\text{e-}4$ and $\lambda_1(2,4) = -1.45\text{e-}4$ (panel A); and $\lambda_1(1,6) = -6.47\text{e-}4$ and $\lambda_1(2,6) = -2.60\text{e-}4$ (panel C). This finding suggests that $\widehat{G}_t$ drives time variations in both expected excess returns to PC_1 and PC_2. In addition, note that the ratio, $\lambda_1(1,\mathcal{N})/\lambda_1(2,\mathcal{N})$, ranges from 2.5 ($\mathcal{N} = 6$) to 6.2 ($\mathcal{N} = 5$), suggesting that $\widehat{G}_t$ influences excess bond returns mainly through its impact on the "level" risk premium.

Second, in all three models $\{CSM(\mathcal{N}\text{-}1,\mathcal{N}),\ \mathcal{N} = 4,5,6\}$, the risk premium driving factors include the first two factors that govern the market prices of "level" and "slope" risks, and the first one appears more important in shaping the unconditional bond risk premia (Koijen et al., 2010). More specifically, persistent contributors to the first risk-premium factor include $PC_{1,t}$, $PC_{2,t}$ and $\widehat{G}_t$; those to the second risk-premium factor include $PC_{3,t}$ and $\widehat{G}_t$. Furthermore, if a model, say, model $CSM(5,6)$, allows for hidden yield factors, then the level risk premium significantly varies with the fifth PC as well (row 1 in panel C). Note that conditioning only on yield curve information (and not on macro variables), the models of Cochrane and Piazzesi (2008) and Duffee (2011) suggest that variations in expected excess bond returns are driven by a single factor.

Third, rows corresponding to $\{PC_{i,t}, i \geq 3\}$ in both λ_0 and λ_1 are uniformly zero in every panel. Hence, among yield PCs only the level and slope risks are priced. This result coincides with

[IA.10] Unreported results indicate that these three properties also emerge in our model selections for unspanned models.

JPS's finding. Duffee (2010) also documents that there are two factors driving the variation in risk premium and presents evidence that this is a robust property of models with the Sharpe ratio constraints. These findings in turn help explain why the restrictions placed on λ_0 and λ_1 make model-implied Sharpe ratios consistent with ones observed in data.

Note that while all three models, $CSM(\mathcal{N}\text{-}1, \mathcal{N})$ with $4 \leq \mathcal{N} \leq 6$, imply non-zero compensation for exposure to the macro risk, the loadings of relevant risk premium on state variables are not robust across models as shown in the last row in each panel. One implication of this result is that the loadings on macro state variables may be difficult to estimate robustly via yield factors in a spanned model. On the other hand, unspanned models are not subject to this problem as λ_0 and λ_1 include no such rows on unspanned macro factors (untabulated).

IA.D Data-Generating Processes Based on VARs

Section 5.2 of the paper presents a finite-sample analysis of Spanning Hypotheses I & II using MTSMs as data-generating processes (DGPs). This section examines an alternative, VAR-based DGP, generated using an approach proposed by Bauer and Hamilton (2018) to address small-sample issues in testing Spanning Hypothesis I (H_0^{S1}).

We first illustrate that the parametric bootstrap design proposed in Bauer and Hamilton (2018; BH hereinafter) is actually more suitable for testing the unconditional predictive power of macro variables than testing H_0^{S1}. We then show that the spanned MTSM specified in Section 5.2.2 provides a more robust test of H_0^{S1} in finite sample analysis than does the VAR-based DGP.

IA.D.1 VAR-based DGPs

BH model the joint dynamics of bond yields and a j-dimensional macroeconomic vector F_t using the following restricted VAR system:

$$Y_t^o = U_{\mathcal{M}} \cdot PC_{1-3,t}^o + \varepsilon_t, \tag{IA.D4}$$

$$\begin{bmatrix} PC_{1-3,t}^o \\ F_t \end{bmatrix} = \begin{bmatrix} \mu_p \\ \mu_f \end{bmatrix} + \begin{bmatrix} \Phi_{pp} & 0_{3\times j} \\ 0_{j\times 3} & \Phi_{ff} \end{bmatrix} \begin{bmatrix} PC_{1-3,t-1}^o \\ F_{t-1} \end{bmatrix} + \begin{bmatrix} \Sigma_p & 0_{3\times j} \\ 0_{j\times 3} & \Sigma_f \end{bmatrix} \begin{bmatrix} \epsilon_t^{\mathcal{P}} \\ \epsilon_t^F \end{bmatrix} \tag{IA.D5}$$

where Y_t^o denotes the time-t observed yields of k zero-coupon bonds with maturities $\mathcal{M} = \{m_1, \ldots, m_k\}$, $U_{\mathcal{M}}$ is a $k \times 3$ matrix with columns equal to the first three eigenvectors of the variance matrix of Y_t^o, and the diagonal matrix ε_t represents fitting errors.

We aim to show that the parameter restrictions specified in Eq. (IA.D5) have a close affinity to the restrictions required for the MTSM in Section 5.1 to satisfy the hypothesis that macro variables have no predictive power for excess bond returns unconditionally (under the $\mathbb{P}$-measure). Following Duffee (2007), we refer to this hypothesis as the "general" null hypothesis (GNH). To proceed, we first introduce such restrictions, termed "macro-independence" restrictions in this study.

IA.D.2 "Macro-Independence" Restrictions

Consider the MTSM in Section 5.1. Given that the expected excess return on an m-period bond from t to $t+j$ is

$$\begin{aligned} E_t\left(rx_{t,t+j}^{(m)}\right) &= \text{constant} + \psi_{m,j}' X_t, \\ \text{where} \quad \psi_{m,j} &= mB_m' - (m-j)B_{m-j}'(\Phi^{\mathbb{P}})^j - jB_j', \end{aligned} \tag{IA.D6}$$

the GNH implies that the last $\mathcal{N}$-$\mathcal{L}$ columns of the model-implied matrix $\psi_{m,j}$ are entirely zero, regardless of bond maturity m or return horizon j. How to implement such restrictions in the model depends on j. Recall that, in our empirical analysis, predictive regressions use annual excess returns sampled at the monthly frequency, while MTSMs are estimated with monthly observations.

Let $\lambda_1 = [\lambda_{1p}, \lambda_{1f}]$ in Eq. (9). If $j = 1$ (month), then setting λ_{1f} to zero prevents macro factors from affecting expected one-period excess returns. Without loss of generality, we allow all $\mathcal{L}$ yield curve factors $\mathcal{P}_t$ to drive variations in bond risk premia. As a result,

$$\psi_{m,1} = -(m-1)B_{m-1}'\lambda_1 \;=\; -(m-1)B_{m-1}'\left[\lambda_{1p}\,,\, 0_{\mathcal{N}\times(\mathcal{N}-\mathcal{L})}\right]. \tag{IA.D7}$$

Under this specification, $E_t\left(rx_{t,t+1}\right)$ is orthogonal to the macro state vector F_t. However, F_t can still affect longer-horizon ($j > 1$) excess returns because future monthly returns, $\{E_t\left(rx_{t+i,t+i+1}\right)\}_{i\geq 1}$, are not orthogonal to F_t. For instance, note that $E_{t+1}\left(rx_{t+1,t+2}\right)$ is determined by $\mathcal{P}_{t+1}$ and $E_t(\mathcal{P}_{t+1})$ depends on F_t. Consequently, F_t contains information about future excess annual returns.

As a result, when $j > 1$, to ensure the state variables determining term premia to vary inde-

pendently of the macro factors, we specify the following $\mathbb{P}$-measure dynamics of X_t:

$$X_t = \begin{bmatrix} \mathcal{P}_t \\ F_t \end{bmatrix} = \begin{bmatrix} \mu_p^{\mathbb{P}} \\ \mu_f^{\mathbb{P}} \end{bmatrix} + \begin{bmatrix} \Phi_{pp}^{\mathbb{P}} & 0_{\mathcal{L}\times(\mathcal{N}-\mathcal{L})} \\ 0_{(\mathcal{N}-\mathcal{L})\times\mathcal{L}} & \Phi_{ff}^{\mathbb{P}} \end{bmatrix} \begin{bmatrix} \mathcal{P}_{t-1} \\ F_{t-1} \end{bmatrix} + \Sigma_x \, \epsilon_{x,t}^{\mathbb{Q}}. \tag{IA.D8}$$

That is, the variation in F_t is independent of expected monthly bond returns at all leads and lags; thus, even for annual excess returns, the last $\mathcal{N}$-$\mathcal{L}$ columns of $\psi_{m,12}$ are constrained to be zero.

Eqs. (IA.D7) and (IA.D8) together lead to the following conditions, termed "macro-independence" restrictions and denoted by H_0^{MI}, for the model to satisfy the GNH:

$$H_0^{MI}: \ \Phi_{fp}^{\mathbb{P}} = 0, \quad \Phi_{pf}^{\mathbb{Q}} = \Phi_{pf}^{\mathbb{P}} = 0, \ \text{ and } \ \Phi_{ff}^{\mathbb{Q}} = \Phi_{ff}^{\mathbb{P}}. \tag{IA.D9}$$

Let $MIM(\mathcal{L},\mathcal{N})$ denote the model subject to these restrictions. Unless specified otherwise, we focus on MTSMs with $\mathcal{N} = \mathcal{L} + 1$ and $F_t = G_t$ in the analysis that follows. For instance, model $MIM(3,4)$ is used below to conduct the finite-sample inference about the GNH.

IA.D.3 VAR-based DGPs and Tests of the GNH

Note that the parameter restrictions specified in Eq. (IA.D5) are very close to the "macro-independence" restrictions given in Eq. (IA.D9) under model $MIM(3,4)$. The only fundamental difference between the VAR-based model in Eqs. (IA.D4) and (IA.D5) and model $MIM(3,4)$ is that the former does not rely on the Duffie and Kan (1996) restrictions for an affine mapping from bond yields to the yield-curve factors. However, empirically this difference is expected to have little impact on the dynamics of expected excess returns, as matrix $U_{\mathcal{M}}$ obtained from the PCA does not significantly deviate from the loading matrix $\mathcal{B}_{\mathcal{M}}$ in Eq. (11).[IA.11] Therefore, like model $MIM(3,4)$, the above VAR-based model implies that term premia are time-varying and driven by yield PCs only; that is, by construction, the macro factors F_t have no predictive power for future yields and bond returns. As such, the VAR-based model in Eqs. (IA.D4) and (IA.D5) satisfies the GNH rather than H_0^{S1} stated in Section 2.2. Put differently, as macro risks are not priced at all in this VAR-based DGP, it is not suitable for conducting tests of evidence for unspanned macro risks.

To further illustrate this point, we generate bootstrap samples using the VAR-based model and

[IA.11] To see this, another equivalent approach to estimating Eq. (11) is regressing the bond yields on yield PCs. While the Duffie-Kan restrictions are not imposed in this estimation (unless the number of factors equals $k-1$), the small magnitude of measurement errors ensures that the OLS-implied loading matrix for $PC_{1-3,t}$ is very close to $\mathcal{B}_{\mathcal{M}}$ if the term structure is truly described by a no-arbitrage dynamic term structure model (Duffee 2010a).

investigate the properties of regression statistics under the same DGP. To proceed, letting F_t be the single SAGLasso factor G_t in the model, we estimate $\mu_p, \mu_f, \Phi_{pp}, \Phi_{ff}, \Sigma_p$, and Σ_f with MLE as in Section 5. Next, we generate bootstrap samples from Eqs. (IA.D4) and (IA.D5) and use a residual bootstrap to resample the PCs and SAGLasso factor based on Eq. (IA.D5). We construct bootstrapped yields, Y_t^b, as follows:

$$Y_t^b = U_{\mathcal{M}} \cdot PC_{1-3,t}^b + \eta_t^b,$$

where $PC_{1-3,t}^b$ denotes the vector of three bootstrapped PCs. Following BH, η_t^b is generated from $MVN(0, \sigma_\eta^2 I)$, where σ_η is set to the sample standard deviation of the fitting errors $\hat{\varepsilon}_t$ (pooled across maturities).[IA.12] Finally, excess bond returns are calculated using bootstrapped yields.

IA.D.4 Finite Sample Analysis Using the VAR-based DGP

What if the above VAR bootstrap design is used to examine the finite-sample properties of the regression in Eq. (1) in tests of H_0^{S1}? To answer this question, we examine finite-sample distributions of regression statistics in testing H_0^{S1}:

$$rx_{t,t+12}^{(12n)} = \alpha + \beta_p' PC_{1-3,t}^o + \beta_g G_t + e_{t+12}. \tag{IA.D10}$$

Table IA.D1 reports the results. A comparison of panels A1–B2 of the table with their counterparts in Table 3 based on model $SM(2,3)$ reveals that the VAR-based bootstrap still understates the size distortions in the regression in Eq. (IA.D10). Indeed, the 5% critical values implied by model $SM(2,3)$ are more than twice as great as those implied by the VAR-based model for most statistics/maturities. The discrepancy between these two DGPs is substantial in both in-sample and out-of-sample analyses and especially glaring in the coefficients of determination. For instance, panel A1 of Table IA.D1 indicates that the upper bound of the 95% confidence interval for ΔR^2 is around 3.3%, but this upper bound is merely comparable to the median of the $SM(2,3)$-implied distributions. More precisely, the VAR-based 5% critical value has a true size of up to 46%, implying that the finite-sample test based on the VAR bootstrap design would reject the null more than eight times as often as it should.

For completeness, panels A3–B4 of Table IA.D1 report the finite-sample distributions implied

[IA.12] We find that replacing these simulated measurement errors with the ones bootstrapped from the actual (maturity-specific) fitting errors has only marginal impact on the finite-sample distributions.

by the macro-independent model $MIM(3,4)$. As expected, they closely resemble their VAR-based counterparts illustrated in panels A1–B2 of the table. Namely, both $MIM(3,4)$ and the VAR-based DGP differ sharply from spanned MTSMs and lead to inflated rejection rates in tests of H_0^{S1}.

To summarize, in our case, finite-sample tests of H_0^{S1} using the VAR-based DGP is actually oversized and thus biased against the null hypothesis. In contrast, the spanned MTSM specified in Section 5.2.2 provides a more relevant and robust test of H_0^{S1} in finite sample analysis.

IA.E Ibragimov-Müller Tests of Spanning Hypotheses I and II

This section conducts an alternative and robust test of H_0^{S1} and H_0^{S2}, drawing an inference about the hypotheses based on the test developed by Ibragimov and Müller (2010; IM hereinafter).

It is known that standard heteroscedasticity and autocorrelation consistent (HAC) corrections perform poorly in small samples. The IM test can improve the performance of these procedures by not relying on consistency of the given variance estimator. In IM's approach, regression coefficients β are estimated q times on q subsets of the whole sample. IM prove that, for each coefficient β_i, the t-statistic computed from the q estimates of $\hat{\beta}_i$ has approximately the same distribution as a standard t-statistic computed from independent and zero-mean Gaussian variables. Müller (2014a) finds that the IM test has outstanding size and power properties in the presence of strongly autocorrelated of regression disturbances. Müller (2014b) further notes that the IM test is an "attractive choice" for predictive regression problem and is also robust to structural breaks.

Following Müller (2014a), we divide the whole sample into q nonoverlapping consecutive blocks of (approximately) equal length, with $q = 8$ or 16. Table IA.E1 reports the p-values of the resultant t-tests of both H_0^{S1} and H_0^{S2}, for both the full and post-1984 samples. As the IM test assumes the independence of blocks, we insert 12-month gaps between adjacent blocks in the full-sample analysis. As such, the regression coefficients estimated from different blocks of data are arguably independent from each other. For brevity, we report the testing results for the average excess bond return only, which is over two- through four-year (ten-year) maturities for the 1964–2014 (1985–2014) sample, as maturity-specific estimates for each of q sample subsets are rather noisy. While the evidence on $PC_{2,t}^o$ (the "slope" factor) is consist with BH, $PC_{1,t}^o$ (the "level" factor) becomes insignificant in the post-1984 sample when $Z_t = \widehat{G}_t$ (the SAGLasso factor). However, even the

strong evidence for the predictive power of $PC^o_{2,t}$ is tempered when we consider H_0^{S2}: its p-value skyrockets to 0.33 and 0.38 in the full and post-1984 samples, respectively. In contrast, the p-values of $\widehat{G}_t$ are uniformly lower than 0.05 for both H_0^{S1} and H_0^{S2}, regardless of the choice of q.

Overall, the IM tests indicate that among the five yield curve factors and the macro factor $\widehat{G}_t$, the latter is the only robust predictor of future excess bond returns at the 5% significance level.

IA.F An Alternative Version of Spanning Hypothesis II

In the tests of H_0^{S2} conducted so far, the yield-curve factors used in the hypothesis are the first five principal components (PCs) of the noise-uncontaminated yield curve. As mentioned in Section 2.2, including the higher-order PCs is motivated by the notion of hidden factors à la Duffee (2011). This section introduces and tests another version of H_0^{S2} that is based on an alternative set of the yield-curve factors, the "cycle" factor ($\widehat{cf}$) of Cieslak and Povala (2015). As noted in Cieslak and Povala (2015), the cycle factor is spanned (see also Cieslak 2018), as well as analogous to the single risk premium factor in Duffee (2011) that contains a hidden component.

Cieslak and Povala (2015) propose an illustrative three-factor dynamic term-structure model (DTSM) in which $\widehat{cf}$ corresponds to a single "risk premium factor" denoted by x_t, where x_t captures all of forecastable variation in one-year expected excess returns for bonds of all maturities. While the Cochrane and Piazzesi (2005) factor ($\widehat{CP}$) plays a similar role in the DTSM proposed in Cochrane and Piazzesi (2008), Cieslak and Povala (2015) demonstrate that their methodology (based on linear projections of yields on trend inflation) is more effective in recovering the variation in risk premiums from noise-contaminated yields and, as a result, $\widehat{cf}$ subsumes $\widehat{CP}$ in predicting excess bond returns. In other words, x_t is analogous to Duffee (2011)'s single risk premium factor, RP_t, that determines the *one-month-ahead* risk premia on all bonds.[IA.13] In particular, x_t contains a hidden component that cannot be detected using the cross-section of yields and that needs to be inferred, say, with a proxy for trend inflation as done in Cieslak and Povala (2015). In this sense, x_t can be regarded as an "annual" version of RP_t and, accordingly, $\widehat{cf}$ maps to the smoothed estimate of RP_t obtained in Duffee (2011). That is, as an estimate of x_t, $\widehat{cf}$ summarizes all information on *one-year-ahead* risk premia.

[IA.13] The state vector underlying the five-factor DTSM in Duffee (2011) consists of the first five PCs of yield innovations. As a result, RP_t is a linear combination of these five PCs.

It follows that we can formulate an alternative version of H_0^{S2} using $\widehat{cf}$ as the conditioning variable:

> $H_0^{S2,cf}$: The SAGLasso macro factor $\widehat{G}$ has no additional predictive power for bond risk premia in the presence of $\widehat{cf}$.

One way to test $H_0^{S2,cf}$ is based on the following predictive regression of excess bond returns:

$$rx_{t,t+12}^{(12n)} = \alpha + \beta_c' \widehat{cf}_t + \beta_g' \widehat{G}_t + e_{t+12}. \tag{IA.F11}$$

As mentioned in Section 4.4.4, we find that in this setting $\widehat{\beta}_g$ is highly significant—based on asymptotic distributions of test statistics. See Table IA.F1 for the results from both in-sample (panel A) and out-of-sample (panel B) tests of $H_0^{S2,cf}$.[IA.14]

To understand finite-sample properties of the regression in Eq. (IA.F11), we extend Cieslak and Povala (2015)'s three-factor DTSM to include the macro factor G_t, and then use this extended model as the DGP for simulation. Note that the structure of this DGP is the same as that presented in Section 5.3.1, except that the state vector here is rotated to $\mathbb{X}_t = (\tau_t, r_t^r, x_t, G_t)$, where τ_t denotes trend inflation and r_t^r the real short rate. Following Cieslak and Povala (2015), we measure τ_t and r_t^r by τ_t^{CPI} and $c_t^{(1)}$, respectively when estimating the model, where τ_t^{CPI} is a discounted moving average of the past 10-year realized core CPI with the gain parameter of 0.99 and $c_t^{(1)}$ the fitted residual from univariate regressions of yields $y_t^{(12n)}$ on τ_t^{CPI}.

The bootstrap design adopted for the regression in Eq. (IA.F11), however, departs from that discussed in Section 5.2 in two aspects. First, the entire model is presented at the annual frequency. As such, we adopt Cieslak and Povala (2015)'s specification of one-period-ahead risk premia:

$$\Sigma\Lambda_t = \begin{bmatrix} \lambda_{0\tau} \\ \lambda_{0\mathfrak{r}} \\ 0_{2\times 1} \end{bmatrix} + \begin{bmatrix} 0 & 0 & \lambda_{1\tau} & 0 \\ 0 & 0 & \lambda_{1\mathfrak{r}} & 0 \\ & & 0_{2\times 4} & \end{bmatrix} \mathbb{X}_t \tag{IA.F12}$$

This specification guarantees that x_t fully determines variations in expected excess *annual* returns. Second, as in Cieslak and Povala (2015), τ_t, r_t^r, and x_t are assumed to evolve independently of each

[IA.14] As is the case of our out-of-sample exercises for the post-1984 sample, as described in Section 4.4.2, we use the initial 15 years as our training sample. That is, the initial coefficient estimates are obtained based on the period from November 1971 to October 1986. Both $\widehat{cf}$ and $\widehat{G}$ are constructed recursively afterwards.

other. It follows that the $\mathbb{P}$-measure dynamics of the state variables are

$$\mathbb{X}_t = \mu_x^{\mathbb{P}} + \begin{bmatrix} \phi_{\tau\tau}^{\mathbb{P}} & 0 & 0 & \phi_{\tau g}^{\mathbb{P}} \\ 0 & \phi_{\mathfrak{r}\mathfrak{r}}^{\mathbb{P}} & 0 & \phi_{\mathfrak{r} g}^{\mathbb{P}} \\ 0 & 0 & \phi_{xx}^{\mathbb{P}} & \phi_{xg}^{\mathbb{P}} \\ \phi_{x\tau}^{\mathbb{P}} & \phi_{g\mathfrak{r}}^{\mathbb{P}} & \phi_{gx}^{\mathbb{P}} & \phi_{gg}^{\mathbb{P}} \end{bmatrix} \mathbb{X}_{t-1} + \begin{bmatrix} \sigma_{\tau\tau} & 0 & 0 & \sigma_{\tau g} \\ 0 & \sigma_{\mathfrak{r}\mathfrak{r}} & 0 & \sigma_{\mathfrak{r} g} \\ 0 & 0 & \sigma_{xx} & \sigma_{xg} \\ \sigma_{x\tau} & \sigma_{g\mathfrak{r}} & \sigma_{gx} & \sigma_{gg} \end{bmatrix} \epsilon_{x,t}^{\mathbb{P}}. \tag{IA.F13}$$

Note that the parameter $\phi_{xg}^{\mathbb{P}}$ is important as it determines whether G_t has unconditional predictive power for excess bond returns; if $\phi_{xg}^{\mathbb{P}} = 0$, the model in Eq. (IA.F13) degenerates into conformity with Duffee (2007)'s GNH. Under the specification in Eq. (IA.F12), G_t contains no *conditional* predictive power when x_t is controlled for, regardless of the value of $\phi_{xg}^{\mathbb{P}}$.

As, by definition, the risk-premium factor x_t does not affect the short rate, we specify the following equation to complete the model:

$$r_t = \delta_0 + \delta_r r_t^r + \delta_\tau \tau_t, \tag{IA.F14}$$

where r_t denotes the one-year yield with $\delta_r > 0$ and $\delta_\tau > 0$.

The MTSM as represented by Eqs. (IA.F12)–(IA.F14) is estimated using zero yields with maturities of one through ten years over the full sample period 1971.11–2014.12 (matching the beginning of the sample used in Cieslak and Povala 2015). The estimated model is then used to generate 5,000 bootstrapped data samples.

Table IA.F1 summarizes the finite-sample properties of the six test statistics used that are based on the 5,000 bootstrapped data samples, including the 95th percentile of the bootstrap distribution (underlined as the 5% critical value in the table) and the p-value (in angle brackets) for each test statistic. A comparison of these finite-sample critical values with those (under H_0^{S2}) reported in panels B3 and B4 of Table 3 reveals that the small-sample bias is less severe in the regression in Eq. (IA.F11) than that in Eq. (1) specified for testing H_0^{S2}. Consistent with the conclusion drawn from their asymptotic distributions, the bootstrap distributions of all statistics shown in Table IA.F1 overwhelmingly reject the null hypothesis $H_0^{S2,cf}$—that $\widehat{G}_t$ contains no predictive power conditioned on $\widehat{cf}_t$—at the 5% significance level, with the only exception of the ENC-REG test for the 7-year bond for which the small-sample p-value is 6.6%.

To summarize, the above results of tests of the spanning hypothesis $H_0^{S2,cf}$ provide further

evidence that the SAGLasso macro factor has significant, additional predictive power for excess bond returns conditioning on the yield curve information.

IA.G Unspanning Tests and Applications of Unspanned Models

This section focuses on unspanned MTSMs with the SAGLasso factor as the sole macro risk factor. We formally describe the macro-unspanning hypothesis (MUH) in Section 5.3.1 and then investigate its statistical significance as well as its economic importance in Sections 5.3.2 through IA.G.2. Lastly, Section IA.G.3 quantifies the information content of the SAGLasso factor.

While the above test results demonstrate the empirical relevance of G_t as an unspanned macro risk, the tests of the MUH per se are more interesting statistically than economically. On the one hand, to be statistically legitimate, the MUH has to be formulated as Eq. (12). On the other hand, the consensus is that, in general, macro variables hold greater promise in helping to improve a term structure model's time-series accuracy than its goodness-of-fit (e.g., Duffee (2011) finds that a $YTSM(5)$ is adequate for producing fitting errors of 6 bps). Given this insight, a more relevant question to ask is whether using G_t as a pricing factor has any economic benefits. Put differently, does an unspanned MTSM with G_t as its sole macro factor provide any added *economic* value over an otherwise spanned model? As shown below, the answer to these questions depends on whether MTSMs are subject to certain constraints on their model-implied Sharpe ratios.

IA.G.1 Model-Implied Sharpe Ratios

One issue not addressed in the likelihood-ratio tests considered in Section 5.3.2 (as well as in BR) is that the MTSMs under scrutiny impose no constraints on the Sharpe ratio (SR) of bond returns and that such "unconstrained" models may imply unrealistic SRs, as noted in Duffee (2010) and Joslin, Singleton, and Zhu (2011; JSZ hereafter). Specifically, Duffee documents that while the empirical benchmark for the unconditional maximum SR is 0.15∼0.20, SRs implied from unconstrained Gaussian dynamic term structure models in his analysis are much higher than the benchmark.

Untabulated results indicate that among the three spanned MTSMs, $\{SM(\mathcal{L},\mathcal{N})\}_{3\leq\mathcal{L}\leq5}$, considered in panel A of Table 4, even the most "reasonable" model-implied sample mean (population

mean) of conditional maximum SRs is 0.715 (0.825) when SRs are computed with log returns; the sample mean increases to 1.309 when SRs are computed with simple returns. Consistent with Duffee (2010), we find that the model-implied SRs increase with the model dimension. For model $SM(5,6)$, the sample mean of the maximum conditional SRs could even be higher than 10^{35} (for simple returns), an obviously unplausible level.[IA.15] As such, though statistically appealing, the test results presented in panel A of Table 4 are based on misspecified models.

One way to ensure that an MTSM generates plausible SRs is to directly impose restrictions on risk premia, say, that only the level and slope risks be priced (a restriction suggested by Duffee 2010 and implemented in Duffee 2011). Another way is to let the data decide what restrictions are empirically relevant. We implement the latter approach by following JPS to search for the best zero restrictions on risk premium parameters $\{\lambda_0, \lambda_1\}$ that minimize the Bayesian information criterion (note that Λ_t essentially represents SRs of bond portfolios with payoffs that track the pricing factors). The resultant models selected by this approach (see Internet Appendix IA.C.2) all possess the following two properties: (a) variations in expected excess bond returns are mainly driven by two factors and, (b) the SAGLasso factor plays a significant role in both term-premium factors. Importantly, conditional maximum SRs implied by these selected models are generally in line with those observed empirically. For convenience, the MTSMs with the selected zero restrictions on $\{\lambda_0, \lambda_1\}$ are referred to as constrained MTSMs and denoted by $CSM(\mathcal{L},\mathcal{N})$ ($CUSM(\mathcal{L},\mathcal{N})$) for spanned (unspanned) models, with $\mathcal{L}$ being the number of yield factors included in the model.

With model selections performed on market prices of risk, unspanned and spanned models are no longer nested, however, and as a result, the LR test-based statistical inference made in Section 5.3.2 no longer applies. Nonetheless, as shown below we can still measure the economic values of the macro-unspanning restrictions imposed on constrained models.

IA.G.2 Out-of-Sample Forecasts of Bond Yields

This subsection investigates whether it is beneficial to include the SAGLasso factor as unspanned by the yield curve in an MTSM. We consider MTSMs with and without the macro-unspanning restrictions and examine whether these restrictions help to forecast future bond yields. We seek to

[IA.15] Untabulated results indicate that this problem persists in MTSMs tested by BR, in which our SAGLasso factor is replaced with (GRO, INF), two macro factors often used in this literature.

quantify the effectiveness of these restrictions as forecasting tools.

We focus on six-factor models in this analysis given Duffee's (2011) argument that five yield factors summarize all information (in both the time series and cross section) in the yield curve. Regardless, including at least five yield factors instead of three makes it harder to see the importance of the macro-unspanning restrictions.

The procedure is similar to the out-of-sample analysis in Section 4.4.2. However, since recursive estimation of MTSMs is computationally very costly (especially for models $CSM(5,6)$ and $CUSM(5,6)$, which require model selection for the risk premium), our yield forecasts are formed based on model estimates for the 1985–2007 sample. With the model parameters fixed, we refilter yield factors using observations up to month t ($\geq$ 2007.12) and then construct forecasts of the T-year bond yield in month-$(t+h)$, where $T = 0.5, 1, 3, 5, 7, 10$ and $h = 1, 3, 6, 12$ in our empirical analysis. The out-of-sample (test) period extends from January 2008 to December 2013.

Panel A of Table IA.G1 reports the root mean squared forecast error (RMSE) produced by unconstrained models $SM(5,6)$ and $USM(5,6)$ for each of 24 combinations of T and h. Note that the models deliver closely comparable forecasting performance, especially at short horizons. This finding is not surprising, given that they should produce identical yield forecasts if $PC_{1-5,t}$ is assumed to be observed without error (see Section 4.2 of JSZ). Although our assumption that all bonds (and portfolios) are priced imperfectly prevents us from exploiting the JSZ-type separation of parameters in the likelihood function, the assumption allows the macro-unspanning restrictions to affect the filtering process and thus the model estimations. As indicated by our empirical results, this impact makes a sizable difference only at the one-year horizon, where $USM(5,6)$ provides more accurate forecasts at the short end of the yield curve but is outperformed by $SM(5,6)$ at the long end. Nonetheless, recall that both $SM(5,6)$ and $USM(5,6)$ generate unrealistic model-implied SRs.

Panel B of Table IA.G1 shows the results from constrained models $CSM(5,6)$ and $CUSM(5,6)$. They too have similar forecasting performance when the forecast horizon is short with h=1,3 (month). However, when h=6 or 12, $CUSM(5,6)$ significantly outperforms $CSM(5,6)$, especially for the 1-year and longer maturity yields. For example, when h=12, the unspanning restrictions reduce the forecast error by as much as 30 bps for the 3-year yield or 40 bps for the 7-year yield. That is, the improvements in forecasting performance owing to an unspanned $\widehat{G}$ are much more robust once certain zero restrictions on Λ_t are imposed. To decipher the discrepancy between these two

pairs of models, we examine the model-implied $\mathbb{P}$-dynamics. As discussed by JPS (in their Section IV.B), enforcing zero restrictions on their risk premium parameters increases the persistence of state variables. We confirm this finding by noting that the eigenvalues of $\Phi_x^{\mathbb{P}}$ in $CUSM(5,6)$ are substantially larger than their counterparts in $CSM(5,6)$. This increase prevents variations in risk premia from completely dominating short-rate expectations and makes model-implied long-dated yield expectations more reasonable and potentially closer to the "true" yield expectations.

Taking the above findings together with the LR test results presented in Section 5.3.2, we conclude that by making the model more parsimonious, the macro-unspanning restrictions do not hurt the in-sample fitting and thus boost the out-of-sample performance.

IA.G.3 Forecastable Variations in Excess Returns Attributable to G_t

Having explored the unspanned nature of G_t, we quantify the information content in G_t within a (G-based) MTSM. Specifically, we examine how much of the predictable variations in excess bond returns can be captured by G_t and the potential role of hidden yield factors in the model. Put differently, we examine how much information related to the bond risk premium may be lost by excluding unspanned macro risks from term structure modeling. Note that this exercise represents an MTSM-based version of the regression analysis conducted in Section IA.B.5.

To this end, we consider the constrained models only (because this exercise requires reasonable model-implied moments of risk premia), and focus on the unspanned models.[IA.16] We implement models $CUSM(\mathcal{L},\mathcal{N})$ for $\mathcal{L}=3,4,5$.

IA.G.3.1 Variance Decomposition for Excess Bond Returns

We discuss the population properties of annual excess bond returns. Results reported in Table IA.G2 cover the five-year bond only as it is closely related to the "in-four-years-for-one-year" forward premium, as shown in the following:

$$E_t\left(rx_{t+12}^{(60)}\right) = FP_t^{4,1} - 4E_t\left(\Delta y_{t+12}^{48}\right) + \left(E_t(y_{t+48}^{(12)}) - y_t^{(12)}\right).$$

But the results for other maturities are broadly similar.

[IA.16] Although the macro-unspanning restrictions tend to grant macro factors the "privilege" of retaining their contributions to term premia, this is less of an issue here given that G_t is constructed after controlling for the yield curve information. Regardless, the spanned models generate qualitatively similar results (untabulated).

Consider the model $CUSM(3,4)$ first. Its model-implied unconditional mean of excess bond returns is 2.85% (column 2), consistent with its data counterpart of 2.73% (untabulated). The unconditional variance is 58.25 (column 3) and calculated using the following formula:

$$\begin{aligned} \text{Var}\left(rx_{t,t+12}^{(60)}\right) &= \psi' \,\text{Var}(X_t)\psi + 48^2 \left[\sum_{i=0}^{11} B_{48}'\Phi^i \Sigma\Sigma' \Phi^{i\prime} B_{48}\right] + (5^2+4^2+1)\sigma_{\eta_y}^2, \quad \text{(IA.G15)} \\ \text{where} \quad \psi &= 60B_{60}' - 48B_{48}'\Phi^{12} - 12B_{12}'. \end{aligned}$$

Among the three terms on the right-hand side (RHS) of Eq. (IA.G15), the first one represents the unconditional variance of the conditional expectation, which quantifies forecastable variation in the excess bond return; the second term denotes the variance of shocks to the "true" excess return; and the last term is the variance of the measurement error's contribution to the observed return shocks. Since this last term is typically small in models with $\mathcal{N} \geq 3$, the predictability of bond returns is mainly determined by the relative magnitudes of the first two terms on the RHS.

Furthermore, how much of $\text{Var}\left(rx_{t,t+12}^{(60)}\right)$ is forecastable depends on the conditioning information used to forecast. If the state vector X_t itself is used, then the full-information R^2 implied by $CUSM(3,4)$ is 0.463 (the ratio of 27.01 in column 4 to 58.25), comparable to the regression R^2 of 0.439 reported in Table IA.B6 (column 7). The full-information R^2, however, cannot be achieved when the conditioning variables consist of the first $\mathcal{R}\,(\leq \mathcal{L})$ PCs of observed yields only. An effective measure for the gap between the information contained in X_t and that in $PC_{1-\mathcal{R},t}^o$ is the following ratio of variances of these two relevant forecasts:

$$VR_{\mathcal{R}}^o = \frac{\psi' \,\text{Var}(X_t|PC_{1-\mathcal{R},t}^o)\psi}{\psi' \,\text{Var}(X_t)\psi}, \quad \mathcal{R} \leq \mathcal{L}. \tag{IA.G16}$$

If $\mathcal{R} = 3$, then $CUSM(3,4)$ implies a $VR_{\mathcal{R}}^o$ of 71.4% (Table IA.G2, column 5 in braces); that is, almost 30% of the information in X_t is lost if we ignore G_t and rely solely on $PC_{1-3,t}^o$ to infer term premia.[IA.17]

What happens if the first $\mathcal{R}\,(\leq \mathcal{L})$ PCs of the *true* yields are used as the conditioning variables? We can repeat the above analysis using the following variant of Eq. (IA.G16):

$$VR_{\mathcal{R}} = \frac{\psi' \,\text{Var}(X_t|PC_{1-\mathcal{R},t})\psi}{\psi' \,\text{Var}(X_t)\psi}. \tag{IA.G17}$$

Column 6 shows that VR_3 is 72.9% (in braces), only slightly greater than VR_3^o (71.4%). This is

[IA.17] Duffee (2011) uses $VR_{\mathcal{R}}^o$ to evaluate the importance of yield factors hidden from the contemporaneous term structure and finds that $PC_{1-3,t}^o$ recover only 70% of the information on expected excess returns on the five-year bond, consistent with the notion of hidden factors.

not surprising as the cross-sectional effect of $PC_{1-3,t}$ is supposedly large enough to dominate the measurement error. Obviously, if the conditioning variables are X_t, the variance ratio is 100% (column 9). Notice that because $\mathcal{R} = \mathcal{L} = 3$, results from $CUSM(3,4)$ in columns 7 and 8 are the same as those in columns 5 and 6.

Given that model $CUSM(3,4)$ leaves no room for hidden yield factors, we consider the higher-dimensional models ($\mathcal{L} > 3$) next. As expected, in such cases the information loss will be higher (than with $\mathcal{L} = 3$) if the conditioning information consists of $PC^o_{1-3,t}$ only. As shown in column 5, VR^o_3 is about 71% under $CUSM(4,5)$ and 65% under $CUSM(5,6)$ (a model that is supposed to encompass both unspanned yield and macro factors). That is, about one-third of the information in X_t is lost if only $PC^o_{1-3,t}$ are used to infer term premia under $CUSM(5,6)$. As before, replacing $PC^o_{1-3,t}$ with $PC_{1-3,t}$ hardly reduces the information lost, with VR_3 equal to 71.3% under $CUSM(4,5)$ and 67.2% under $CUSM(5,6)$ (column 6). Note from column 7 that including higher-order PCs of the *observed* yield curve, $PC^o_{4,t}$ and $PC^o_{5,t}$, in the conditioning variables hardly helps to dig up more information on risk premia. For instance, for $CUSM(5,6)$, $VR^o_5 = 65.1\%$ (column 7), only slightly higher than $VR^o_3 = 64.9\%$ (column 5). Again, this is because the cross-sectional effect of higher-order PCs is too small to overwhelm the measurement error.

If we can *perfectly* infer the hidden factors by extracting information from yield dynamics as well as in the cross section, we can estimate risk premia more accurately. For instance, under model $CUSM(5,6)$, $VR_5 = 75.8\%$ (column 8), much higher than either $VR^0_5 = 65.1\%$ (column 7) or $VR_3 = 67.2\%$ (column 6). In fact, this difference between VR^o_5 and VR_5 suggests a wedge between the information in observed yields and that in "true" yields, whereas there is no evidence for a similar gap for the first three PCs, as indicated by columns 5 and 6. Nonetheless, the VR_5 of 75.8% still implies an information loss of almost 25% even in this ideal case. Given that under model $CUSM(5,6)$, G_t is not spanned by $PC_{1-6,t}$ and that the five yield factors presumably summarize all (time-series and cross-sectional) information on the yield side (Duffee 2011), a more reasonable implication of the result that $VR_5 = 75.8\%$ is the following: The information loss is *at least* about one-quarter when G_t is excluded from return predictors, even though they include $PC_{1-5,t}$.

We use the phrase "at least" for two reasons: First, the variance ratio is computed under the assumption that $PC_{1-5,t}$ are perfectly observable. In practice, however, econometricians have to perform filtering analysis to infer $PC_{4-5,t}$. Duffee (2011) documents that the Kalman filter recovers

only two-thirds of the information in the true state vector for monthly excess returns (and about 82% of that for annual excess returns in an earlier version of the paper). In contrast, measurement error has little impact on factor G_t. Second, the period 1985–2007 sample is special in the sense that the fraction of the total variance attributable to macro-driven variations is particularly low. If the estimation sample is extended to either 1964 or 2014, the model-implied variance ratio would drop below 67% (untabulated). Once these two facts are taken into account, the results from model-based risk premium decomposition are expected to be close to the test results of H_0^{S2} (columns 14–17 in Table IA.B6)—namely, with respect to the state vector $X_t = (PC_{1-5,t}, G_t)$, the SAGLasso factor accounts for almost half of the predictable variations in excess bond returns.

It is worth emphasizing that risk premium accounting based on variance ratios is analogous to the variance decomposition (in the context of reduced-form VARs), of which the results are sensitive to the order of state factors chosen for identification. The projection of G_t on $PC_{1-\mathcal{L},t}$ in $VR_{\mathcal{L}}$ maximizes the explanatory power of yield PCs (Bikbov and Chernov 2010). This point can be illustrated by changing the order of state factors and calculating the following variance ratio:

$$VR_{3+G} = \frac{\psi' \operatorname{Var}(X_t | X_t^{\setminus H})\psi}{\psi' \operatorname{Var}(X_t)\psi}, \tag{IA.G18}$$

where $X_t^{\setminus H} = (PC_{1-3,t}, G_t)$. Results in column 9 indicate that under $CUSM(5,6)$, the first three PCs plus the SAGLasso factor capture 97.9% of forecastable variation in excess bond returns. Although this finding does not necessarily mean that hidden factors are unimportant in return prediction in this case, it does imply that, compared to ignoring hidden yield factors (as shown in column 9), excluding unspanned macro risks (associated with G_t, as shown in column 8) bears more serious economic consequences in the inference of term premia.

IA.G.3.2 Calculations of Variance Ratios

This subsection provides details on the calculations of variance ratios used in Section IA.G.3.1. All the calculations are based on MTSMs specified in either Section 5.2 (for spanned models) or Section IA.G (for unspanned ones).

Consider $VR^o_{\mathcal{R}}$, the variance ratio defined in Eq. (IA.G16) that focuses on the forecast of excess bond returns based on the first $\mathcal{R}\,(\leq \mathcal{L})$ PCs of observed yields. Recall that by definition, the first

$\mathcal{L}$ PCs are given by

$$PC^o_{1-\mathcal{L},t} = W_{\mathcal{R},\mathcal{M}} Y^o_t = W_{\mathcal{R},\mathcal{M}} \mathcal{A}_{\mathcal{M}} + W_{\mathcal{R},\mathcal{M}} \mathcal{B}'_{\mathcal{M}} X_t + W_{\mathcal{R},\mathcal{M}}\, \eta_t,$$

where $W_{\mathcal{R},\mathcal{M}}$ is an $\mathcal{R} \times k$ loading matrix, which is equal to the transpose of $U_{\mathcal{R},\mathcal{M}}$ in Eq. (IA.D4). Below subscripts are suppressed for simplicity of notations. It follows that the expectation of the true state factor X_t conditioned on these PCs equals

$$E(X_t|PC^o_{1-\mathcal{R},t}) = E(X_t) + \mathrm{Var}(X_t)\mathcal{B}W' \,\mathrm{Var}(PC^o_{1-\mathcal{R},t})^{-1} PC^o_{1-\mathcal{R},t},$$

where the variance of $PC^o_{1-\mathcal{R},t}$ is

$$\mathrm{Var}(PC^o_{1-\mathcal{R},t}) = W\mathcal{B}' \,\mathrm{Var}(X_t)\mathcal{B}W + WW'\sigma^2_\eta.$$

The variance of $E(X_t|PC^o_{1-\mathcal{R},t})$ is

$$\mathrm{Var}(X_t|PC^o_{1-\mathcal{R},t}) = \mathrm{Var}(X_t)\mathcal{B}W' \,\mathrm{Var}(PC^o_{1-\mathcal{R},t})^{-1} W\mathcal{B}' \,\mathrm{Var}(X_t).$$

Next, consider $VR_{\mathcal{R}}$, the variance ratio specified in Eq. (IA.G17) that concerns the inference of risk premium based on the first $\mathcal{R}$ PCs of true yields. Recall that these PCs, $PC_{1-\mathcal{R},t}$, constitute a segment of the state vector X_t. Denoting the remaining $\mathcal{N}$-$\mathcal{R}$ state factors by $X_t^{\setminus\mathcal{R}}$, we have

$$\begin{aligned} E(X_t|PC_{1-\mathcal{R},t}) &= \begin{bmatrix} PC_{1-\mathcal{R},t} \\ E(X_t^{\setminus\mathcal{R}}) + \mathfrak{C}\mathfrak{V}^{-1} PC_{1-\mathcal{R},t} \end{bmatrix}, \quad \text{and} \\ \mathrm{Var}(X_t|PC_{1-\mathcal{R},t}) &= \begin{bmatrix} \mathfrak{V} & \mathfrak{C}' \\ \mathfrak{C} & \mathfrak{C}\mathfrak{V}^{-1}\mathfrak{C}' \end{bmatrix}, \end{aligned}$$

where $\mathfrak{V} = \mathrm{Var}(PC_{1-\mathcal{R},t})$ and $\mathfrak{C} = \mathrm{Cov}(X_t^{\setminus\mathcal{R}}, PC_{1-\mathcal{R},t})$.

Table IA.A1: Properties of Principal Components of Observed Yield Curves

Panel A reports correlations between principal components (PCs) of observed yields and filtered estimates of yield PCs, denoted by PC_i^o and PC_i, respectively, where $i = 1, \ldots, 5$. Population correlations are computed by simulating 100,000 months of bond yields. The 95% confidence intervals for the sample correlations, as displayed in parentheses, are derived from 5,000 simulations with the same number of observations as in the data sample. The yield maturities in all simulations are three months and one through five years. Panel B reports results from regressions of the return to an n-year zero-coupon bond from month t to month $t+12$ less the month-t yield on a one-year bond on the first five PCs of observed yields, $PC_{1-5}^o = (PC_1^o, \ldots, PC_5^o)$. Test statistics are computed using either the Hansen and Hodrick (1980) GMM covariance estimator (in parentheses) or the Newey and West (1987) HAC covariance estimator (in brackets). The row labeled "R^2 (Table 2)" copies the R^2 values from regressions of excess returns on filtered estimates of the first five PCs, reported in Table IA.B5 (columns 10 through 13). The ΔR^2 measure represents the differences between R^2 values in Panel B and R^2 (Table 2). The last row in panel B reports the percentage decrease in the R^2. The sample spans the period January 1964 to December 2014.

(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Panel A: Corr($PC_{i,t}$,$PC_{i,t}^o$)		Panel B: Predictive regressions of excess returns to an n-year zero-coupon bond on $PC_{1-5,t}^o$			
			Bond maturity n			
	Population	Sample	2	3	4	5
$PC_{1,t}^o$	0.9999	0.9998	3.526	3.031	2.080	0.422
		[0.9998 0.9999]	(1.116)	(0.529)	(0.264)	(0.043)
			[1.266]	[0.601]	[0.300]	[0.048]
$PC_{2,t}^o$	0.9905	0.9902	-0.688	-1.330	-2.038	-2.621
		[0.9885 0.9916]	(-3.650)	(-3.674)	(-3.961)	(-4.105)
			[-4.021]	[-4.092]	[-4.431]	[-4.601]
$PC_{3,t}^o$	0.9612	0.9818	0.784	1.011	1.485	1.688
		[0.9787 0.9845]	(1.233)	(0.920)	(1.041)	(0.978)
			[1.384]	[1.034]	[1.156]	[1.079]
$PC_{4,t}^o$	0.7233	0.7595	-1.956	-2.836	-2.798	-0.961
		[0.7238 0.7912]	(-1.702)	(-1.295)	(-0.910)	(-0.244)
			[-1.842]	[-1.418]	[-1.005]	[-0.271]
$PC_{5,t}^o$	0.2125	0.6107	4.060	10.521	15.004	14.677
		[0.5584 0.6581]	(2.693)	(4.536)	(5.636)	(4.358)
			[2.363]	[3.620]	[4.074]	[3.151]
		R^2	0.205	0.215	0.241	0.228
		R^2 (Table IA.B5)	0.221	0.233	0.255	0.245
		ΔR^2	-0.016	-0.018	-0.014	-0.017
		Percentage decrease in R^2	-7.24%	-7.73%	-5.49%	-6.94%

Table IA.B1: Correlation between Yield Curve and New Macro Factors

This table reports the Pearson correlation coefficients between four newly constructed macroeconomic factors and five yield-curve factors. The four macroeconomic factors include employment ($\hat{g}_{1t}$), housing ($\hat{g}_{2t}$), inflation ($\hat{g}_{3t}$), and the aggregate SAGLasso factor ($\widehat{G}_t$) constructed in Section 4.2. The five yield curve factors include the first three principal components (PCs) of *observed* bond yields, $\{PC^o_{i,t}, i = 1, 2, 3\}$, and the filtered higher-order PCs of noise-uncontaminated yields, $PC_{4,t}$ and $PC_{5,t}$. The sample spans the period January 1964 to December 2014.

	$\widehat{G}_t$	$\widehat{G}_{1t}$	$\widehat{G}_{2t}$	$\widehat{G}_{3t}$	$PC^o_{1,t}$	$PC^o_{2,t}$	$PC^o_{3,t}$	$PC_{4,t}$
$\widehat{G}_{1t}$	0.620							
$\widehat{G}_{2t}$	0.527	0.577						
$\widehat{G}_{3t}$	0.524	0.467	0.351					
$PC^o_{1,t}$	−0.100	−0.226	−0.167	−0.199				
$PC^o_{2,t}$	−0.352	−0.222	−0.073	−0.270	−0.006			
$PC^o_{3,t}$	0.167	0.239	0.222	0.196	0.018	0.003		
$PC_{4,t}$	−0.094	−0.031	−0.106	0.021	−0.000	0.013	0.044	
$PC_{5,t}$	−0.021	−0.027	−0.282	0.284	0.024	0.008	−0.011	0.092

Table IA.B2: Predictive Power of Three SAGLasso Group Factors

The return to an n-year zero-coupon bond from month t to month $t+12$ less the month-t yield on a one-year bond is regressed on $\widehat{g}_{1t}$, $\widehat{g}_{2t}$, and $\widehat{g}_{3t}$, the group factors, constructed in Section 4.2, that represent employment, housing and inflation, respectively, for $n = 2, \ldots, 5$. Results reported in panel A are based on the January 1964–December 2014 sample, including those from both univariate (panel A1) and multivariate predictive regressions (panel A2). Results reported in panel B are based on the January 1952–December 2014 sample, where only the employment factor is considered (because of data limitations) and constructed using this longer sample ($\widehat{g}^*_{1t}$). In-sample results from regressions on $\widehat{g}^*_{1t}$ are shown in panel B1 and out-of-sample (OOS) results based on $\widetilde{g}^*_{1t}$ are reported in panel B2. Test statistics are computed using either the Hansen and Hodrick (1980) GMM covariance estimator (in parentheses) or the Newey and West (1987) HAC covariance estimator (in brackets). The ENC-REG statistic denotes the OOS t-statistic proposed by Ericsson (1992), whose 95th percentile of the asymptotic distribution is $\Phi^{-1} = 1.645$. The row labeled "ENC-NEW" reports a variant of the ENC-REG statistic proposed by Clark and McCracken (2001); their simulation shows that the 95% critical value is around 1.584 for testing one additional predictor. Both tests share the same null hypothesis that the benchmark model encompasses the unrestricted model with excess predictors. The R^2_{oos} statistic denotes the OOS R^2 of Campbell and Thompson (2008).

maturity n (year)	2	3	4	5		2	3	4	5
					Panel A: Sample period 1964–2014				
	Panel A1: Univariate Regressions					Panel A2: Multivariate Regressions			
$\widehat{g}_{1t}$	0.828	1.526	2.082	2.568		0.942	1.401	1.744	2.062
	(3.796)	(4.149)	(4.272)	(4.545)		(2.850)	(3.040)	(3.038)	(3.165)
	[4.237]	[4.594]	[4.689]	[4.948]		[3.107]	[3.301]	[3.280]	[3.401]
R^2	0.220	0.222	0.213	0.211					
$\widehat{g}_{2t}$	0.643	1.162	1.631	2.051		0.722	0.968	1.223	1.461
	(2.608)	(2.861)	(3.178)	(3.364)		(1.965)	(2.052)	(2.215)	(2.321)
	[2.969]	[3.249]	[3.589]	[3.786]		[2.141]	[2.236]	[2.413]	[2.528]
R^2	0.143	0.139	0.141	0.149					
$\widehat{g}_{3t}$	0.723	1.358	1.872	2.222		0.847	1.249	1.574	1.763
	(3.096)	(3.075)	(3.023)	(2.875)		(2.544)	(2.534)	(2.486)	(2.358)
	[3.449]	[3.431]	[3.380]	[3.215]		[2.754]	[2.748]	[2.698]	[2.555]
R^2	0.168	0.176	0.172	0.173		0.404	0.431	0.420	0.417
					Panel B: Sample period 1952–2014				
	Panel B1: In-Sample Regressions					Panel B2: $\widetilde{g}^*_{1t}$ vs. constant (OOS)			
$\widehat{g}^*_{1t}$	0.751	1.397	1.932	2.390					
	(4.161)	(4.524)	(4.767)	(5.084)	ENC-REG	3.329	3.600	3.719	3.991
	[4.665]	[5.025]	[5.227]	[5.525]	ENC-NEW	136.93	134.56	127.99	125.341
R^2	0.206	0.213	0.211	0.211	R^2_{oos}	0.155	0.164	0.166	0.169

Table IA.B3: Unspanned Variation in SAGLasso Group Factors

This table reports results from linear projections of each of the three SAGLasso group macro factors $\{\widehat{g}_{it}, 1 \leq i \leq 3\}$ onto the first $\mathcal{R}$ principal components (PCs) of the yield curve ($PC^o_{1-\mathcal{R},t}$), where $\mathcal{R} = 3$ (panel A) or 6 (panel B) and the group factors are the employment ($\widehat{g}_{1t}$), housing ($\widehat{g}_{2t}$), and inflation ($\widehat{g}_{3t}$) factors. Columns 3 and 5 show the regression R^2s, and in brackets beneath are reported 95% confidence intervals based on 5,000 artificial samples simulated from a six-factor constrained term structure model with spanned macro risks. The state vector of the model, denoted by $CSM(3,6)_{group}$ and specified in Section IA.G.1, includes three yield curve factors (the first three PCs) and three macro factors, $\widehat{g}_{1t}$, $\widehat{g}_{2t}$, and $\widehat{g}_{3t}$. Column 2 indicates whether the three macro variables are assumed to be measured with errors in the estimation of model $CSM(3,6)_{group}$. Columns 4 and 6 report the first-order serial correlation of regression residuals.

(1)	(2)	(3)	(4)	(5)	(6)
Dependent variable	Macro Measure-ment Error	Panel A: Regressions of $\widehat{g}_{it}$ on $PC^o_{1-3,t}$		Panel B: Regressions of $\widehat{g}_{it}$ on $PC^o_{1-6,t}$	
		R^2	AR(1) of residuals	R^2	AR(1) of residuals
$\widehat{g}_{1t}$		0.116		0.125	
	No	[0.148 0.770]		[0.206 0.809]	
	Yes	[0.140 0.773]	0.951	[0.201 0.814]	0.948
$\widehat{g}_{2t}$		0.082		0.119	
	No	[0.144 0.816]		[0.171 0.845]	
	Yes	[0.137 0.802]	0.960	[0.164 0.829]	0.946
$\widehat{g}_{3t}$		0.106		0.123	
	No	[0.152 0.708]		[0.219 0.766]	
	Yes	[0.145 0.697]	0.903	[0.216 0.753]	0.893

Table IA.B4: Predictive Power of Alternative Macroeconomic Factors for Excess Bond Returns

The return to an n-year zero-coupon bond from month t to month $t+12$ less the month-t yield on a one-year bond is regressed on a macro factor (F_t), for $n = 2, \ldots, 5$. The three macro factors considered include the modified Ludvigson and Ng (2011) factor, $\widehat{LN}_t^m$ (panel A); a factor denoted $\widehat{G}_t^{rev}$ (panel B) and constructed using the SAGLasso procedure (Section 4.1) albeit with the set of 131 macro series that are not adjusted for data revisions; and a factor denoted $\widehat{G}_t^{rev,lag}$ (panel C) and constructed using the SAGLasso procedure albeit with the set of 131 macro series that are not adjusted for either data revisions or publication lags. In the in-sample analysis, t-statistics are computed using the Hansen and Hodrick (1980) GMM covariance estimator (in parentheses) and the Newey and West (1987) HAC covariance estimator (in brackets), respectively. In the out-of-sample analysis, the row labeled "ENC-REG" reports the out-of-sample t-statistics proposed by Ericsson (1992), whose 95th percentile of the asymptotic distribution is $\Phi^{-1} = 1.645$. The row labeled "ENC-NEW" reports a variant of the ENC-REG statistic proposed by Clark and McCracken (2001); their simulation shows that the 95% critical value is around 1.584 for testing one additional predictor. Both tests share the same null hypothesis that the benchmark model encompasses the unrestricted model with excess predictors. The row labeled "R^2_{oos}" denotes the out-of-sample R^2 of Campbell and Thompson (2008). The sample spans the period January 1964 to December 2014.

	Predictive regressions of excess returns to an n-year zero-coupon bond on alternative macro factors (F_t)											
	Panel A: $F_t = \widehat{LN}_t^m$				Panel B: $F_t = \widehat{G}_t^{rev}$				Panel C: $F_t = \widehat{G}_t^{rev,lag}$			
maturity n (year)	2	3	4	5	2	3	4	5	2	3	4	5
	Panel A1: In-sample				Panel B1: In-sample				Panel C1: In-sample			
Coeff. on F_t	0.724	1.473	2.181	2.792	1.039	1.993	2.832	3.542	1.135	2.150	3.037	3.807
	(4.227)	(4.932)	(5.804)	(6.554)	(5.780)	(6.357)	(6.715)	(7.009)	(6.023)	(6.412)	(6.681)	(7.102)
	[4.641]	[5.390]	[6.319]	[7.074]	[6.215]	[6.802]	[7.191]	[7.513]	[6.547]	[6.946]	[7.183]	[7.595]
R^2	0.168	0.207	0.234	0.250	0.347	0.379	0.394	0.402	0.414	0.441	0.453	0.464
	Panel A2: Out-of-sample				Panel B2: Out-of-sample				Panel C2: Out-of-sample			
ENC-REG	2.660	3.336	4.136	5.104	3.867	4.975	5.933	7.266	4.308	5.217	5.647	6.558
ENC-NEW	115.96	137.06	162.59	173.71	159.78	182.11	207.94	219.67	204.64	235.34	253.81	262.21
R^2_{oos}	0.051	0.112	0.186	0.225	0.118	0.200	0.274	0.314	0.123	0.238	0.294	0.326

Table IA.B5: In-Sample Tests of Spanning Hypotheses I and II: 1964–2014

The return to an n-year zero-coupon bond from month t to month $t+12$ less the month-t yield on a one-year bond is regressed respectively on (i) the first three principal components (PCs) of observed bond yields $PC^o_{1-3,t}$ (columns 2-5); (ii) $PC^o_{1-3,t}$ and the SAGLasso macro factor $\widehat{G}_t$ (columns 6-9); (iii) the filtered first five PCs of noise-uncontaminated yields $PC_{1-5,t}$ (columns 10-13); and (iv) $PC_{1-5,t}$ and $\widehat{G}_t$ (columns 14-17). Test statistics are computed using the Hansen and Hodrick (1980) GMM covariance estimator (in parentheses), or the Newey and West (1987) HAC covariance estimator (in brackets). The sample spans the period January 1964 to December 2014.

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)
					Spanning Hypothesis I								Spanning Hypothesis II			
maturity n (year)	2	3	4	5	2	3	4	5	2	3	4	5	2	3	4	5
$\mathcal{P}1_t$	3.636	3.286	2.426	0.738	5.512	6.774	7.176	6.535	2.193	2.163	1.918	1.216	3.034	3.713	4.020	3.777
	(1.106)	(0.533)	(0.285)	(0.070)	(2.810)	(1.814)	(1.364)	(1.010)	(1.280)	(0.692)	(0.443)	(0.226)	(3.091)	(2.058)	(1.527)	(1.147)
	[1.257]	[0.608]	[0.325]	[0.079]	[3.097]	[2.014]	[1.521]	[1.126]	[1.451]	[0.786]	[0.504]	[0.257]	[3.378]	[2.272]	[1.704]	[1.280]
$\mathcal{P}2_t$	−0.688	−1.328	−2.036	−2.619	−0.236	−0.489	−0.893	−1.224	−0.441	−0.838	−1.279	−1.636	−0.151	−0.303	−0.553	−0.752
	(−3.351)	(−3.291)	(−3.559)	(−3.789)	(−1.338)	(−1.413)	(−1.808)	(−1.994)	(−3.720)	(−3.657)	(−3.924)	(−4.071)	(−1.424)	(−1.477)	(−1.900)	(−2.048)
	[−3.732]	[−3.709]	[−4.028]	[−4.288]	[−1.456]	[−1.565]	[−2.019]	[−2.231]	[−4.124]	[−4.098]	[−4.413]	[−4.580]	[−1.554]	[−1.638]	[−2.116]	[−2.284]
$\mathcal{P}3_t$	0.782	1.000	1.466	1.665	0.118	−0.233	−0.213	−0.385	0.480	0.571	0.811	0.857	0.053	−0.217	−0.258	−0.444
	(1.174)	(0.849)	(0.958)	(0.917)	(0.214)	(−0.236)	(−0.166)	(−0.258)	(1.065)	(0.742)	(0.799)	(0.694)	(0.150)	(−0.372)	(−0.336)	(−0.474)
	[1.319]	[0.956]	[1.065]	[1.015]	[0.236]	[−0.260]	[−0.179]	[−0.274]	[1.167]	[0.812]	[0.866]	[0.749]	[0.160]	[−0.394]	[−0.349]	[−0.486]
$\mathcal{P}4_t$									−0.766	−2.467	−4.004	−5.112	−0.205	−1.431	−2.600	−3.402
									(−0.988)	(−1.880)	(−2.396)	(−2.384)	(−0.312)	(−1.338)	(−1.925)	(−1.982)
									[−0.999]	[−1.871]	[−2.337]	[−2.330]	[−0.331]	[−1.397]	[−1.960]	[−2.007]
$\mathcal{P}5_t$									4.375	9.333	12.097	11.581	4.424	9.423	12.220	11.730
									(2.185)	(2.573)	(2.405)	(1.858)	(2.967)	(3.534)	(3.322)	(2.540)
									[2.241]	[2.629]	[2.472]	[1.914]	[3.104]	[3.691]	[3.489]	[2.681]
$\widehat{G}_t$					1.007	1.872	2.548	3.110					1.005	1.853	2.513	3.061
					(5.958)	(6.370)	(6.649)	(6.542)					(6.575)	(7.100)	(7.135)	(6.892)
					[6.268]	[6.599]	[6.847]	[6.809]					[7.042]	[7.573]	[7.628]	[7.401]
R^2	0.167	0.156	0.183	0.194	0.422	0.418	0.432	0.436	0.221	0.233	0.255	0.245	0.472	0.486	0.494	0.477
ΔR^2 due to $\widehat{G}_t$					0.255	0.262	0.250	0.243					0.251	0.253	0.240	0.232

Table IA.B6: In-Sample Tests of Spanning Hypotheses I and II: 1985–2014

The return to an n-year zero-coupon bond from month t to month $t+12$ less the month-t yield on a one-year bond is regressed respectively on (i) the first three principal components (PCs) of observed bond yields $PC^o_{1-3,t}$ (columns 2-5); (ii) $PC^o_{1-3,t}$ and the SAGLasso macro factor $\widehat{G}_t$ (columns 6-9); (iii) the filtered first five PCs of noise-uncontaminated yields $PC_{1-5,t}$ (columns 10-13); and (iv) $PC_{1-5,t}$ and $\widehat{G}_t$ (columns 14-17). Test statistics are computed using the Hansen and Hodrick (1980) GMM covariance estimator (in parentheses), or the Newey and West (1987) HAC covariance estimator (in brackets). The sample spans the period January 1985 to December 2014.

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)
					Spanning Hypothesis I								Spanning Hypothesis II			
maturity n (year)	2	5	7	10	2	5	7	10	2	5	7	10	2	5	7	10
$\mathcal{P}^o_{1,t}(\mathcal{P}_{1,t})$	6.268	11.951	15.082	19.997	6.160	11.544	14.522	19.305	2.877	7.345	10.466	14.210	2.482	5.863	8.469	11.699
	(2.172)	(1.168)	(0.979)	(1.001)	(3.629)	(2.066)	(1.564)	(1.511)	(2.385)	(1.701)	(1.618)	(1.631)	(3.939)	(2.794)	(2.270)	(2.193)
	[2.439]	[1.314]	[1.101]	[1.119]	[3.881]	[2.224]	[1.720]	[1.672]	[2.668]	[1.915]	[1.827]	[1.834]	[4.058]	[2.909]	[2.491]	[2.435]
$\mathcal{P}^o_{2,t}(\mathcal{P}_{2,t})$	−0.274	−1.450	−2.255	−3.210	−0.039	−0.568	−1.043	−1.711	−0.107	−0.709	−1.281	−1.662	0.003	−0.294	−0.722	−0.960
	(−1.279)	(−2.013)	(−2.331)	(−2.850)	(−0.223)	(−1.024)	(−1.466)	(−2.194)	(−1.269)	(−2.369)	(−3.020)	(−3.101)	(0.041)	(−1.247)	(−2.188)	(−2.532)
	[−1.468]	[−2.292]	[−2.626]	[−3.179]	[−0.253]	[−1.160]	[−1.635]	[−2.410]	[−1.416]	[−2.646]	[−3.359]	[−3.407]	[0.046]	[−1.408]	[−2.463]	[−2.819]
$\mathcal{P}^o_{3,t}(\mathcal{P}_{3,t})$	0.187	−0.525	−1.867	−2.318	0.244	−0.311	−1.573	−1.954	−0.022	−0.295	−1.253	−0.885	0.003	−0.203	−1.129	−0.729
	(0.524)	(−0.439)	(−1.146)	(−1.054)	(1.096)	(−0.396)	(−1.457)	(−1.445)	(−0.100)	(−0.428)	(−1.483)	(−0.783)	(0.016)	(−0.356)	(−1.626)	(−0.952)
	[0.561]	[−0.469]	[−1.203]	[−1.123]	[1.215]	[−0.446]	[−1.601]	[−1.625]	[−0.103]	[−0.439]	[−1.481]	[−0.776]	[0.017]	[−0.389]	[−1.725]	[−1.031]
$\mathcal{P}_{4,t}$									−0.222	−0.305	−0.388	−0.297	−0.152	−0.041	−0.033	0.150
									(−0.757)	(−0.313)	(−0.291)	(−0.183)	(−0.647)	(−0.054)	(−0.031)	(0.122)
									[−0.835]	[−0.345]	[−0.320]	[−0.201]	[−0.711]	[−0.059]	[−0.034]	[0.134]
$\mathcal{P}_{5,t}$									1.023	2.230	0.790	0.356	0.870	1.657	0.018	−0.615
									(2.101)	(1.573)	(0.396)	(0.136)	(2.038)	(1.456)	(0.011)	(−0.308)
									[2.164]	[1.451]	[0.359]	[0.122]	[2.199]	[1.391]	[0.010]	[−0.280]
$\widehat{G}_t$					0.713	2.677	3.679	4.553					0.693	2.605	3.512	4.414
					(6.419)	(6.330)	(5.576)	(5.221)					(6.480)	(6.225)	(5.551)	(5.278)
					[6.822]	[6.841]	[6.014]	[5.636]					[6.819]	[6.674]	[5.945]	[5.676]
R^2	0.140	0.142	0.170	0.206	0.393	0.439	0.448	0.471	0.184	0.176	0.225	0.240	0.417	0.450	0.472	0.483
ΔR^2 due to $\widehat{G}_t$					0.253	0.297	0.278	0.265					0.232	0.274	0.247	0.243

Table IA.B7: Tests of Spanning Hypotheses Using Macroeconomic Variables with Different Lags

This table reports both the in-sample and out-of-sample tests of spanning hypotheses, using macroeconomic variables with 0, 3, 6, 9, or 12 lags. The benchmark model in tests of Spanning Hypothesis I is based on the first three principal components (PCs) of observed bond yields and, in tests of Spanning Hypothesis II, is based on the filtered first five PCs of noise-uncontaminated yields. The columns labeled "HH" report test statistics based on the Hansen and Hodrick (1980) GMM covariance estimator, and the columns labeled "NW" report their counterparts based on the Newey and West (1987) HAC covariance estimator. "ENC-REG" denotes the out-of-sample t-statistics proposed by Ericsson (1992), and "ENC-NEW" represents a variant of the ENC-REG statistic proposed by Clark and McCracken (2001). The rows labeled "ΔR^2" (ΔR^2_{oos}) represent the incremental in-sample (out-of-sample) R^2 due to $\tilde{G}$. The sample spans the period January 1964 to December 2014.

Lags	In-sample tests						Out-of-sample tests					
	2-year bond			5-year bond			2-year bond			5-year bond		
	HH	NW	ΔR^2	HH	NW	ΔR^2	ENC-REG	ENC-NEW	ΔR^2_{oos}	ENC-REG	ENC-NEW	ΔR^2_{oos}
Panel A: Tests of Spanning Hypotheses I												
0	6.711	7.021	0.240	6.136	6.394	0.194	5.168	151.150	0.299	4.548	103.117	0.205
3	6.425	6.845	0.280	6.414	6.716	0.239	4.570	221.703	0.331	3.761	156.355	0.228
6	5.958	6.268	0.255	6.542	6.809	0.243	4.764	191.912	0.349	4.871	147.103	0.271
9	7.449	7.634	0.289	7.104	7.033	0.225	3.481	163.430	0.211	2.935	98.113	0.104
12	6.947	7.367	0.305	6.795	7.170	0.244	5.420	236.068	0.325	4.606	151.385	0.206
Panel B: Tests of Spanning Hypotheses II												
0	7.392	7.918	0.226	6.758	7.220	0.176	4.928	135.506	0.296	4.389	83.981	0.186
3	7.059	7.617	0.265	7.004	7.455	0.223	5.286	243.125	0.387	3.861	150.602	0.239
6	6.575	7.042	0.251	6.892	7.401	0.232	4.781	180.940	0.353	4.526	130.102	0.256
9	8.931	9.081	0.274	7.982	8.048	0.210	3.656	152.386	0.204	2.835	82.606	0.073
12	8.444	8.863	0.289	7.748	8.314	0.227	4.916	230.903	0.330	3.717	136.416	0.186

Table IA.C1: Estimates of Parameters on the Market Price of Risk

This table reports the maximum likelihood estimates of parameters λ_0 and λ_1 that govern bond risk premia in an $\mathcal{N}$-factor constrained, spanned macro-finance term structure model (MTSM) as specified and denoted $CSM(\mathcal{L},\mathcal{N})$ in Section IA.G.1. The underlying state variables include the SAGLasso macro factor, $\widehat{G}$, constructed in Section 4.1 and the first $\mathcal{L}$ principal components (PCs) of bond yields, $PC_{1-\mathcal{L}} = (PC_1, \ldots, PC_{\mathcal{L}})$. The three MTSMs considered include $CSM(3,4)$ (panel A), $CSM(4,5)$ (panel B), and $CSM(5,6)$ (panel C). The one-period risk premium is as specified in Eq. (13): $\Sigma\Lambda_t = \lambda_0 + \lambda_1 X_t = \lambda_0 + \lambda_1 \cdot (PC_{1-\mathcal{L},t},\ \widehat{G}_t)'$. Zero entries of λ_0 and λ_1 reflect our model selection outcome. Values in parentheses are standard errors computed using Monte Carlo simulations.

State variables	λ_0	λ_1 $(\mathcal{N}\times\mathcal{N})$					
		$\lambda_1(\cdot,\mathcal{N})$	$\lambda_1(\cdot,1)$		$\cdots$		$\lambda_1(\cdot,5)$
		$\widehat{G}_t$	$PC_{1,t}$	$PC_{2,t}$	$PC_{3,t}$	$PC_{4,t}$	$PC_{5,t}$
Panel A: Model $CSM(3,4)$							
$PC_{1,t}$	0.013	-6.11e-04	-0.054	-0.313	0		
	(0.002)	(8.97e-05)	(0.016)	(0.087)			
$PC_{2,t}$	0.002	-1.45e-04	0	0	-0.458		
	(9.31e-04)	(7.12e-05)			(0.0143)		
$PC_{3,t}$	0	0	0	0	0		
$\widehat{G}_t$	-0.278	-0.159	0.093	0	0		
	(0.152)	(0.081)	(0.035)				
Panel B: Model $CSM(4,5)$							
$PC_{1,t}$	0.018	-8.13e-04	-0.049	-0.152	0	0	
	(0.003)	(9.04e-05)	(0.007)	(0.060)			
$PC_{2,t}$	0.002	-1.32e-04	0	-0.031	-0.129	0.140	
	(0.001)	(9.13e-05)		(0.035)	(0.139)	(0.148)	
$PC_{3,t}$	0	0	0	0	0	0	
$PC_{4,t}$	0	0	0	0	0	0	
$\widehat{G}_t$	0	0	-0.633	0	0	-8.77	
			(0.243)			(4.871)	
Panel C: Model $CSM(5,6)$							
$PC_{1,t}$	0.029	-6.47e-04	-0.048	-0.173	0	0	-0.708
	(0.003)	(8.76e-05)	(0.009)	(0.076)			(0.259)
$PC_{2,t}$	0	-2.60e-04	0	0	-0.207	0.098	0
		(9.26e-05)			(0.102)	(0.115)	
$PC_{3,t}$	0	0	0	0	0	0	0
$PC_{4,t}$	0	0	0	0	0	0	0
$PC_{5,t}$	0	0	0	0	0	0	0
$\widehat{G}_t$	-0.646	0	0	0	0	0	0
	(0.084)						

Table IA.D1: Finite-Sample Properties of Statistics in Testing Spanning Hypothesis I under a VAR-based Data-Generating Process

This table presents results based on finite-sample distributions of the statistics that are involved in tests of Spanning Hypotheses I stated in Section 2.2. The analysis is based on 5,000 bootstrapped samples generated from the reduced-form VAR described in Eqs. (IA.D4) and (IA.D5) (panels A1 through B2) or from the macro-finance term structure model $MIM(3,4)$ (panels A3 through B4) that satisfies the "macro-independence restrictions" given in Eq. (IA.D9). The length of each bootstrapped sample is set to be consistent with either the full sample (panel A) or the post-1984 subsample (panel B). Test statistics considered include those computed using the Hansen and Hodrick (1980) GMM covariance estimator (HH), the Newey and West (1987) HAC covariance estimator (NW) with 18 lags, the out-of-sample ENC-REG test of Ericsson (1992), or the out-of-sample ENC-NEW test of Clark and McCracken (2001). For each test statistics, the 95th percentile of the bootstrap distribution is reported as the 5% critical value, and the p-values (in angle brackets) are the frequency of bootstrap replications in which the test statistics are at least as large as the statistic in the data. The "ΔR^2" and "ΔR^2_{oos}" measures denote the incremental R^2 and out-of-sample R^2 of Campbell and Thompson (2008), respectively.

	Panel A: Full sample, 1964–2014				Panel B: Subsample, 1985–2014			
maturity (year)	2	3	4	5	2	5	7	10
	Panel A1: In-sample based on VAR				Panel B1: In-sample based on VAR			
HH	1.872	1.889	1.883	1.885	1.999	1.965	2.013	2.055
	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩
NW	1.984	1.997	1.986	1.997	2.031	2.015	2.039	2.059
	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩
ΔR^2	0.032	0.033	0.033	0.034	0.036	0.037	0.037	0.035
	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩
	Panel A2: Out-of-sample based on VAR				Panel B2: Out-of-sample based on VAR			
ENC-REG	1.784	1.747	1.752	1.753	2.002	2.102	2.019	2.013
	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.008⟩	⟨0.001⟩	⟨0.001⟩	⟨0.001⟩
ENC-NEW	10.711	11.286	11.676	11.950	6.688	6.758	6.604	6.506
	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩
ΔR^2_{oos}	0.031	0.031	0.029	0.029	0.047	0.048	0.048	0.047
	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩
	Panel A3: In-sample based on $MIM(3,4)$				Panel B3: In-sample based on $MIM(3,4)$			
HH	1.888	1.921	1.912	1.939	1.994	2.001	2.030	2.003
	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩
NW	1.999	2.021	2.043	2.041	2.046	2.027	2.041	2.039
	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩
ΔR^2	0.032	0.033	0.033	0.033	0.039	0.039	0.038	0.039
	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩
	Panel A4: Out-of-sample based on $MIM(3,4)$				Panel B4: Out-of-sample based on $MIM(3,4)$			
ENC-REG	1.749	1.716	1.729	1.756	1.870	1.995	2.057	2.033
	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.007⟩	⟨0.002⟩	⟨0.003⟩	⟨0.001⟩
ENC-NEW	11.236	11.217	11.101	11.424	6.313	6.326	6.361	6.313
	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩
ΔR^2_{oos}	0.030	0.030	0.030	0.030	0.048	0.051	0.053	0.052
	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩

Table IA.E1: Ibragimov-Müller Test of Spanning Hypotheses I and II

The average return to zero-coupon bonds from month t to month $t+12$ less the month-t yield on a one-year bond is regressed on either $PC^o_{1-3,t}$ and $\widehat{G}_t$ for Spanning Hypothesis I (H_0^{S1}) or $PC_{1-5,t}$ and G_t for Spanning Hypothesis II (H_0^{S2}), where $PC^o_{1-3,t}$ denotes the first three principal components (PCs) of observed bond yields; $PC_{1-5,t}$ the filtered first five PCs of noise-uncontaminated yields; and $\widehat{G}_t$ the SAGLasso single factor. All reported quantities are the p-values for the Ibragimov-Müller (2010) test of the individual significance of the coefficients. The dependent variable is the excess return averaged over 2- through 5-year (10-year) bond maturities in regressions over the full sample period 1964–2014 (the post-1984 subsample). In the full-sample analysis, each block is constructed such that they are 12 months apart from each other.

	Spanning Hypotheses Tested							
	H_0^{S1}		H_0^{S2}		H_0^{S1}		H_0^{S2}	
	Full sample, 1964–2014				Subsample, 1985–2014			
q (# of blocks)	$q=8$	$q=16$	$q=8$	$q=16$	$q=8$	$q=16$	$q=8$	$q=16$
$PC^o_{1,t}(PC_{1,t})$	0.039	0.001	0.006	0.003	0.219	0.417	0.001	0.001
$PC^o_{2,t}(PC_{2,t})$	0.016	0.009	0.327	0.047	0.020	0.006	0.036	0.379
$PC^o_{3,t}(PC_{3,t})$	0.162	0.354	0.309	0.615	0.037	0.647	0.536	0.743
$PC_{4,t}$			0.186	0.961			0.278	0.942
$PC_{5,t}$			0.170	0.107			0.002	0.002
$\widehat{G}_t$	0.009	0.018	0.004	0.014	0.044	0.049	0.015	0.019

Table IA.F1: Tests of An Alternative Version of Spanning Hypotheses II

This table presents asymptotic and finite-sample results from tests of an alternative version of Spanning Hypothesis II, denoted $H_0^{S2,cf}$, that states that the SAGLasso macro factor (Section 4.2) has no additional predictive power for future excess bond returns, conditional on the "cycle" factor of Cieslak and Povala (2015). The tests of $H_0^{S2,cf}$ are based on the following regression, as specified in Eq. (IA.F11):

$$rx_{t,t+12}^{(12n)} = \alpha + \beta_c' \widehat{cf}_t + \beta_g' \widehat{G}_t + e_{t+12},$$

where $rx_{t,t+12}^{(12n)}$ is the excess return to an n-year zero-coupon bond from month t to month $t+12$, for $n = 2, 5, 7, 10$; $\widehat{cf}_t$ denotes the cycle factor; and $\widehat{G}_t$ the SAGLasso macro factor. For the in-sample results (panel A), t-statistics are computed using either the Hansen and Hodrick (1980) GMM covariance estimator (in parentheses) or the Newey and West (1987) HAC covariance estimator (in brackets). Out-of-sample tests considered (panel B) include the "ENC-REG" test of Ericsson (1992) and the "ENC-NEW" test proposed by Clark and McCracken (2001). The ΔR^2 and 'ΔR^2_{oos} measures denote the incremental R^2 and out-of-sample R^2 of Campbell and Thompson (2008), respectively, due to augmenting univariate regressions of $rx_{t,t+12}^{(12n)}$ on $\widehat{cf}_t$ with $\widehat{G}_t$ as in the above equation. The sample spans the period November 1971–December 2014. To obtain the finite-sample distributions of the aforementioned six statistics, 5,000 bootstrapped samples are generated from the term structure model specified in Eqs. (IA.F12)–(IA.F14) in Section IA.F. For each set of test statistics, the 95th percentile of the bootstrap distribution is reported and underlined as the 5% critical value, and the p-values (in angle brackets) are the frequency of bootstrap replications in which the test statistics are at least as large (small) as the statistic in the data.

maturity (year)	2	5	7	10		2	5	7	10
	Panel A: In-sample under $H_0^{S2,cf}$					Panel B: Out-of-sample under $H_0^{S2,cf}$			
$\widehat{G}_t$	0.688	2.491	3.376	4.153					
HH	(4.257)	(4.994)	(4.537)	(4.367)	ENC-REG	1.917	3.099	3.620	4.513
	(1.964)	(2.121)	(2.601)	(2.890)		(1.617)	(2.675)	(3.837)	(4.067)
	⟨0.001⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩		⟨0.033⟩	⟨0.026⟩	⟨0.066⟩	⟨0.029⟩
NW	[4.684]	[5.506]	[4.953]	[4.735]	ENC-NEW	41.321	77.762	80.203	77.102
	[1.872]	[2.120]	[2.588]	[2.857]		[1.374]	[4.429]	[9.080]	[10.746]
	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩		⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩
ΔR^2	0.132	0.144	0.131	0.124	ΔR^2_{oos}	0.053	0.143	0.132	0.103
	(0.011)	(0.032)	(0.062)	(0.072)		(0.010)	(0.033)	(0.065)	(0.075)
	⟨0.000⟩	⟨0.000⟩	⟨0.000⟩	⟨0.001⟩		⟨0.000⟩	⟨0.000⟩	⟨0.001⟩	⟨0.009⟩

Table IA.G1: Out-of-sample Forecasting Performance of Macro-Finance Term Structure Models

This table reports the root mean square errors (RMSEs) for out-of-sample yield forecasts formed from four different macro-finance term structure models (MTSMs). Forecasted yields include the 0.5-, 1-, 3-, 5-, 7-, and 10-year yields. The forecast horizons considered are 1, 3, 6, and 12 months. The four MTSMs—denoted by $SM(\mathcal{L},\mathcal{N})$, $USM(\mathcal{L},\mathcal{N})$, $CSM(\mathcal{L},\mathcal{N})$, and $CUSM(\mathcal{L},\mathcal{N})$—are all driven by six factors ($\mathcal{N}$=6), including five yield curve factors ($\mathcal{L}$=5) and the sole SAGLasso macro factor ($\widehat{G}_t$). Among the four models, $SM(5,6)$ and $USM(5,6)$ are unconstrained models (panel A), and $CSM(5,6)$ and $CUSM(5,6)$ are constrained models (panel B), in which selected zero restrictions are placed on parameters governing bond risk premia. Additionally, while models $SM(5,6)$ and $CSM(5,6)$ are spanned, the other two are unspanned models that satisfy the macro-unspanning conditions (H_0^{US}) specified in Eq. (12). Each of the four MTSMs is estimated once using the 1985-2007 sample. The Kalman filter is implemented recursively with observations from 1985:1 to the time that the forecast is made, beginning in 2008:1 and extending through 2014:12. Errors are reported in basis points of annualized yields.

Panel A: Out-of-sample RMSEs for unconstrained models

Forecast horizon	$y^{(6)}$		$y^{(12)}$		$y^{(36)}$		$y^{(60)}$		$y^{(84)}$		$y^{(120)}$	
	$SM(5,6)$	$USM(5,6)$	$SM(5,6)$	$USM(5,6)$	$SM(5,6)$	$USM(5,6)$	$SM(5,6)$	$USM(5,6)$	$SM(5,6)$	$USM(5,6)$	$SM(5,6)$	$USM(5,6)$
1	25.40	24.32	21.08	20.71	25.62	25.72	27.50	26.79	30.05	29.26	32.29	31.66
3	48.19	46.31	48.30	45.96	54.07	54.98	53.50	54.37	51.95	50.65	48.77	51.97
6	80.27	74.93	81.44	76.26	84.47	85.86	81.61	84.27	76.98	78.20	69.34	73.90
12	139.51	128.08	142.04	130.69	133.32	128.10	114.04	112.76	99.78	99.73	82.13	89.99

Panel B: Out-of-sample RMSEs for constrained models

Forecast horizon	$y^{(6)}$		$y^{(12)}$		$y^{(36)}$		$y^{(60)}$		$y^{(84)}$		$y^{(120)}$	
	$CSM(5,6)$	$CUSM(5,6)$	$CSM(5,6)$	$CUSM(5,6)$	$CSM(5,6)$	$CUSM(5,6)$	$CSM(5,6)$	$CUSM(5,6)$	$CSM(5,6)$	$CUSM(5,6)$	$CSM(5,6)$	$CUSM(5,6)$
1	24.15	27.73	20.00	20.63	25.01	24.86	26.08	26.08	28.71	43.32	30.54	30.63
3	45.77	46.85	45.89	43.75	51.84	48.78	50.92	49.41	48.77	52.60	46.23	46.15
6	74.83	69.99	74.63	63.74	79.10	68.63	79.22	72.83	74.17	63.69	67.11	59.98
12	102.13	95.51	104.02	81.35	116.20	87.04	115.87	91.30	109.84	69.34	92.13	66.01

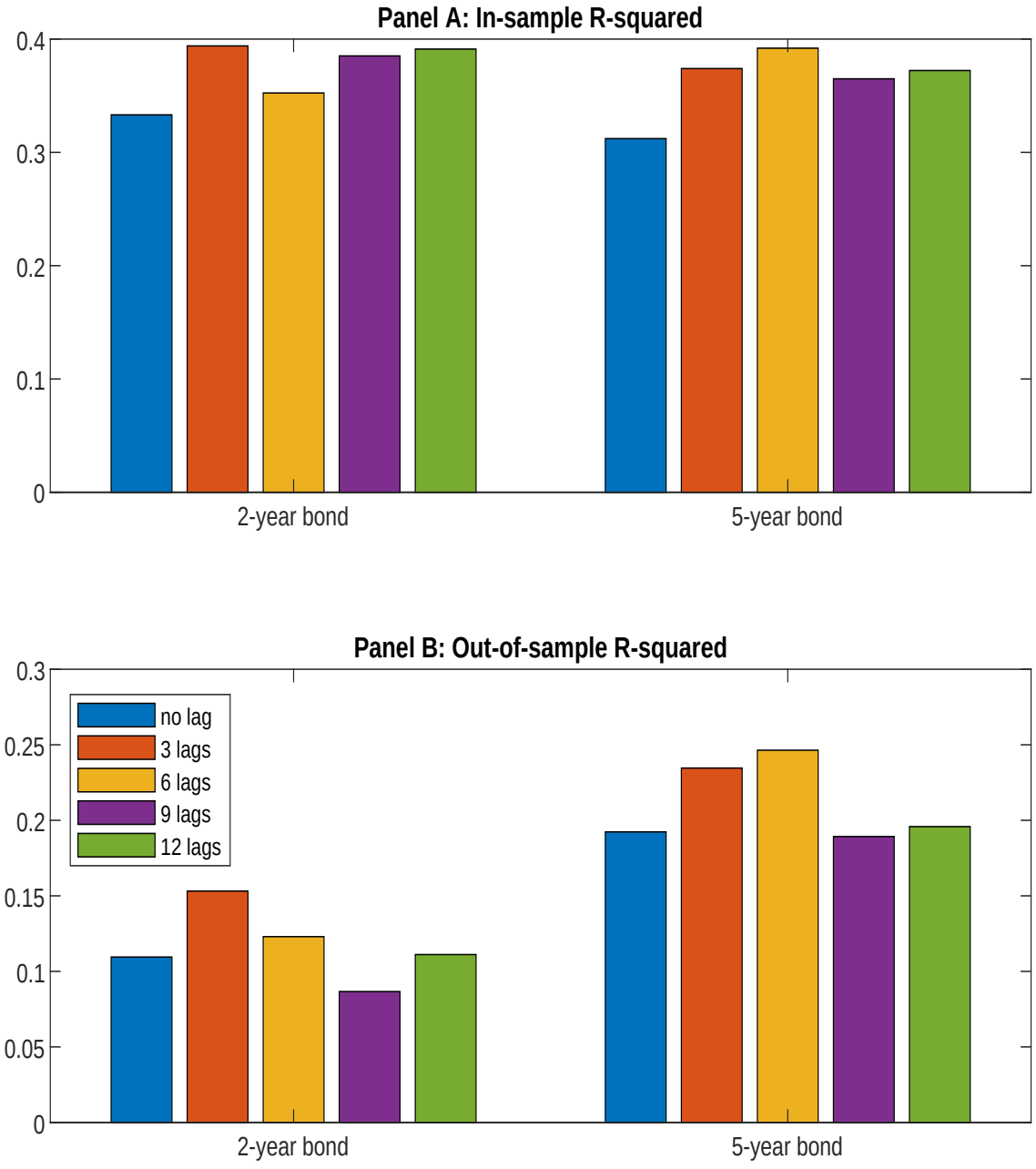
Table IA.G2: Properties of Annual Excess Returns for Five-Year Bonds Implied by Term Structure Models with Unspanned Macro Risks

This table presents the model-implied population moments of unconditional and conditional excess returns on a five-year bond, based on the macro-finance term structure model $CUSM(\mathcal{L},\mathcal{N})$ specified in Section IA.G.1. Model $CUSM(\mathcal{L},\mathcal{N})$ is an $\mathcal{N}$-factor model with unspanned macro risks and "zero restrictions" imposed on risk premium parameters, whose underlying state vector $X_t = (PC_{1-\mathcal{L},t}, G_t)$, where $PC_{1-\mathcal{L},t}$ represent the first $\mathcal{L}$ principal components (PCs) of the *true* yields and G_t the (unspanned) SAGLasso macro factor. The last six columns quantify the variance of true conditional expected excess returns attributable to time variation in the true state vector X_t (column 4), the first three PCs of *observed* yields (column 5), the first three PCs of *true* yields (uncontaminated by measurement errors) (column 6), the first $\mathcal{L}$ PCs of *observed* yields (column 7), the first $\mathcal{L}$ PCs of *true* yields (column 8), and the first three yield PCs plus the SAGLasso factor G_t (column 9), respectively. For each of the last five columns, their ratios to the full-information variance (column 4)—the variance ratios "*VR*"—are reported in braces. The R^2 reported for each of the last three columns is their ratios to the total variance (column 3). The sample period extends from January 1985 to December 2007.

(1)	(2)	(3)		(4)	(5)	(6)	(7)	(8)	(9)
$\mathcal{L}$	Mean	Total Variance		Variance of conditional expectation based on					
				Full info	$PC^o_{1-3,t}$	$PC_{1-3,t}$	$PC^o_{1-\mathcal{L},t}$	$PC_{1-\mathcal{L},t}$	$PC_{1-3,t}$+G_t
3	2.85	58.25		27.01	19.26	19.70	19.26	19.70	27.01
			VR		{0.714}	{0.729}	{0.714}	{0.729}	{1.000}
			R^2				33.1%	33.8%	46.4%
4	2.53	63.13		30.34	21.47	21.62	22.41	23.29	29.76
			VR		{0.708}	{0.713}	{0.739}	{0.768}	{0.965}
			R^2				35.5%	36.9%	47.2%
5	2.50	67.37		33.05	21.45	22.21	21.50	25.05	32.36
			VR		{0.649}	{0.672}	{0.651}	{0.758}	{0.979}
			R^2				31.9%	37.2%	48.1%

Figure IA.B1: Predictive R^2 of Macroeconomic Factors Based on Different Lags

This figure depicts the in-sample and out-of-sample R^2 from bond return predictions with single macroeconomic factors. Macroeconomic factors are constructed from 131 macro variables, along with 0, 3, 6, 9, or 12 of their lags. The sample spans the period January 1964 to December 2014.



References

Barillas, F. 2012. Can we exploit predictability in bond markets? *Available at SSRN 1787567*. Working paper, Emory University.

Bauer, M. D., and J. D. Hamilton. 2018. Robust bond risk premia. *Review of Financial Studies* 31(2):399–448.

Bianchi, D., M. Büchner, and A. Tamoni. 2021. Bond Risk Premia with Machine Learning. *Review of Financial Studies* 34(2):1046–1089.

Bikbov, R., and M. Chernov. 2010. No-arbitrage macroeconomic determinants of the yield curve. *Journal of Econometrics* 159(1):166–182.

Campbell, J., and S. Thompson. 2008. Predicting excess stock returns out of sample: Can anything beat the historical average? *Review of Financial Studies* 21(4):1509–1531.

Cieslak, A. 2018. Short-Rate Expectations and Unexpected Returns in Treasury Bonds. *Review of Financial Studies* 31(9):3265–3306.

Cieslak, A., and P. Povala. 2015. Expected returns in Treasury bonds. *Review of Financial Studies* 28(10):2859–2901.

Clark, T., and M. McCracken. 2001. Tests of equal forecast accuracy and encompassing for nested models. *Journal of Econometrics* 105(1):85–110.

Cochrane, J., and M. Piazzesi. 2005. Bond risk premia. *American Economic Review* 95(1):138–160.

Cochrane, J., and M. Piazzesi. 2008. Decomposing the Yield Curve. *Working Paper, University of Chicago*.

Duffee, G. R. 2007. Are variations in term premia related to the macroeconomy? *Working paper, Johns Hopkins University*.

Duffee, G. R. 2010a. Forecasting with the term structure: The role of no-arbitrage restrictions. *Working paper, Johns Hopkins University*.

Duffee, G. R. 2010b. Sharpe ratios in term structure models. *Working paper, Johns Hopkins University*.

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

Duffee, G. R. 2013. Bond pricing and the macroeconomy. In G. M. Constantinides, M. Harris, and R. M. Stulz (Eds.), *Handbook of the Economics of Finance*, Volume 2B: Asset Pricing, pp. 907–968. North Holland.

Duffie, D., and R. Kan. 1996. A yield-factor model of interest rates. *Mathematical Finance* 6:379–406.

Ericsson, N. 1992. Parameter constancy, mean square forecast errors, and measuring forecast performance: An exposition, extensions, and illustration. *Journal of Policy Modeling* 14(4):465–495.

Ghysels, E., C. Horan, and E. Moench. 2018. Forecasting through the rear-view mirror: Data revisions and bond return predictability. *Review of Financial Studies* 31(2):678–714.

Hansen, L., and R. Hodrick. 1980. Forward exchange rates as optimal predictors of future spot rates: An econometric analysis. *Journal of Political Economy* 88(5):829–853.

Ibragimov, R., and U. K. Müller. 2010. t-Statistic based correlation and heterogeneity robust inference. *Journal of Business & Economic Statistics* 28(4):453–468.

Joslin, S., A. Le, and K. J. Singleton. 2013. Why Gaussian macro-finance term structure models are (nearly) unconstrained factor-VARs. *Journal of Financial Economics* 109(3):604–622.

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

Joslin, S., K. J. Singleton, and H. Zhu. 2011. A new perspective on Gaussian dynamic term structure models. *Review of Financial Studies* 24(3):926–970.

Koijen, R. S. J., T. E. Nijman, and B. J. M. Werker. 2010. When Can Life Cycle Investors Benefit from Time-Varying Bond Risk Premia? *Review of Financial Studies* 23(2):741–780.

Ludvigson, S., and S. Ng. 2009. Macro factors in bond risk premia. *Review of Financial Studies* 22(12):5027–5067.

Ludvigson, S., and S. Ng. 2011. A factor analysis of bond risk premia. In A. Ullah and D. E. A. Giles (Eds.), *Handbook of Empirical Economics and Finance*, pp. 313–372. CRC Press.

Müller, U. K. 2014a. HAC corrections for strongly autocorrelated time series. *Journal of Business & Economic Statistics* 32(3):311–322.

Müller, U. K. 2014b. Rejoinder. *Journal of Business & Economic Statistics* 32(3):338–340.

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