Erasmus University Rotterdam
Erasmus School of Economics
Bachelor Thesis BSc^2 in Econometrics and Economics

Resolving the Spanning Controversy in Macro-Finance Literature

Jean-Frans J. Bart (524132)

Supervisor:	M. Grith
Second assessor:	G. Freire
Date final version:	2nd July 2023

Abstract

In this paper, I investigate whether variables beyond the first three principal components of bond yields can improve the prediction of excess bond returns. By using a bootstrap procedure and considering recent data, I robustly test the incremental predictive power of predictors proposed in earlier literature as well as predictors obtained from a large macroeconomic data set using factor extraction methods, including self-constructed machine-learning methods, both in-sample and out-of-sample. I find that the incremental forecasting power of the predictors proposed by the revisited studies is much weaker than originally suggested. The in-sample predictive power beyond the three yield principal components of the predictor factors resulting from most novel methods is significant. When winsorization is applied, I find that the in-and-out-of-sample incremental predictive power of the predictor factors of one method is consistently significant in all considered sample periods.

The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics or Erasmus University Rotterdam.

1 Introduction

The macro-finance literature presents mixed results concerning the hypothesis that all information relevant for predicting future bond returns is spanned by the yield curve. This "spanning hypothesis" is a central issue in macro-finance (Gürkaynak & Wright, 2012) and has been actively tested over the past years. If the hypothesis holds, predicting future bond returns and estimating bond risk premia would be enormously simplified. Reliable estimates for these risk premia are of great interest to policy-makers, practitioners, and researchers, due to their important implications for monetary policy, investment strategies, and finance theory. Furthermore, since the spanning hypothesis is implied by most macro-finance models, the practical relevance of these models is also implicitly tested when the spanning hypothesis is empirically tested.

In this paper, I investigate whether macroeconomic information can improve bond return forecasts conditional on the first three principal components (PCs) of bond yields. I thus implicitly test a specific version of the spanning hypothesis in which the first three yield PCs are assumed to summarize the information in the yield curve. This assumption seems reasonable as the first three PCs of the bond yields capture almost all the cross-sectional variation in the yields (Litterman & Scheinkman, 1991). I consider both specific macroeconomic variables and factors extracted from a large macroeconomic data set as potential additional predictors for bond returns beyond the three yield PCs. For the extraction of the factors, I use several methods, including self-constructed machine-learning methods. The in-and-out-of-sample incremental predictive power for excess bond returns over the three yield PCs of the different (sets of) additional predictors are evaluated.

Several influential papers provide evidence suggesting that the spanning hypothesis can be rejected. Ludvigson and Ng (2009) show that PCs summarizing a large set of macroeconomic variables are useful for forecasting bond returns, even when controlling for a factor that captures the information in the yield curve. In addition, specific macroeconomic variables have been found to have predictive power beyond the information spanned by the yield curve. In particular, these macroeconomic variables are measures of the output gap (Cooper & Priestley, 2008), supply of Treasury bonds (Greenwood & Vayanos, 2014), inflation and economic activity (Joslin, Priebsch & Singleton, 2014), and trend inflation (Cieslak & Povala, 2015). These findings suggest that there are macroeconomic variables that are not spanned by the yield curve but that are helpful for predicting bond returns. Moreover, Cochrane and Piazzesi (2005) show that, even though the fourth and fifth PCs of the bond yields capture only a very small fraction of the cross-sectional variation in the yields, they are still relevant for forecasting bond returns.

However, the evidence in these papers comes from predictive regressions in which future

excess bond returns are regressed on both yield curve factors, typically the three yield PCs, and additional predictors. The spanning hypothesis is then rejected if the coefficients of the additional predictors in the predictive regression are estimated to be (jointly) significant. Bauer and Hamilton (2018) argue that these regressions have problematic small-sample features that result in too small standard errors for the additional predictors when using conventional tests and thus can lead to a spurious rejection of the spanning hypothesis. To address these issues, Bauer and Hamilton (2018) propose a novel bootstrap procedure specifically designed to avoid this "standard error bias" and thus to correctly test the spanning hypothesis. Using this bootstrap method to estimate the predictive regressions, Bauer and Hamilton (2018) conclude that the in-sample evidence on the rejection of the spanning hypothesis is much weaker than the six influential papers cited above suggested. Moreover, when using new data to evaluate the true out-of-sample forecasts resulting from a restricted model with only the three yield PCs and an unrestricted model with both the three yield PCs and the additional predictors, they find that according to the Diebold and Mariano (2002) test the unrestricted model never leads to significantly better forecasts and mostly even leads to worse out-of-sample forecasts.

Ludvigson and Ng (2009) use the well-known principal component analysis (PCA) method to obtain predictor factors from a large set of macroeconomic variables. However, a drawback of using this method for forecasting is that it completely ignores the information in the target variable. To overcome this weakness, D. Huang, Jiang, Li, Tong and Zhou (2022) propose the scaled principal component analysis (sPCA) method, which improves the PCA method by putting more weight on the variables with stronger predictive power for the target variable. Applying both methods to macroeconomic data, they show that the sPCA method generally performs better in terms of forecasting than the PCA method. D. Huang, Jiang, Li, Tong and Zhou (2023) use both methods to forecast excess bond returns and find that the sPCA method leads to better in-sample and out-of-sample results.

J. Z. Huang and Shi (2023) provide a potential resolution to the spanning controversy. They propose a new two-step machine learning algorithm, called the Supervised Adaptive Group LASSO (SAGLasso) method, to construct a new bond return predictor from a large data set that contains 131 macroeconomic variables along with their six lagged values. They test the spanning hypothesis in two ways. Firstly, they perform out-of-sample tests in which they examine whether the return forecasting model containing both the three yield PCs and the SAGLasso factor as predictors outperforms the restricted model which only contains the three yield PCs. In this out-of-sample analysis, they use the out-of-sample R^2 (Campbell & Thompson, 2008) and two encompassing tests for nested models to evaluate the computed forecasts. Secondly, they use the

framework of Joslin et al. (2014) for macro-finance term structure models to test the hypothesis. From these test outcomes, they conclude that the predictive power of the constructed factor for government bond returns is significant and robust to bond yields.

Based on the intuition of the sPCA method and the method used in J. Z. Huang and Shi (2023) to obtain the SAGLasso factor, I construct various new machine-learning methods to extract predictor factors from a large macroeconomic data set. The predictive power for bond returns beyond the three yield PCs of these factors as well as the factors resulting from the PCA, sPCA and some established machine-learning methods are evaluated both in-sample and out-of-sample using the methodological framework of Bauer and Hamilton (2018). To reduce the impact of extreme values, I also implement winsorization approaches in the factor extraction procedures. In addition, I revisit the evidence in the studies of Ludvigson and Ng (2009), Joslin et al. (2014), Cieslak and Povala (2015), and Cochrane and Piazzesi (2005) using the same data and methodology as Bauer and Hamilton (2018).

After revisiting the evidence in the four published studies with more robust methods for testing, I draw the same conclusions as Bauer and Hamilton (2018). Furthermore, I find that the in-sample predictive power for excess bond returns beyond the three yield PCs of the predictor factors resulting from most factor extraction methods is highly significant. When winsorization is applied, the addition of the predictor factors resulting from most methods to a restricted forecasting model with only the three yield PCs improves the out-of-sample forecasts for excess bond returns. Even though the PCA method is outperformed by other methods in some cases, its in-and-out-of-sample performance is the most consistent and produces predictor factors that have significant predictive power for excess bond returns conditional on the three yield PCs.

I extend the studies of Ludvigson and Ng (2009) and Bauer and Hamilton (2018) by using other methods in addition to the PCA method to extract potential bond return predictors from a large macroeconomic data set, adding new and lagged data to the data set, and implementing winsorization techniques to mitigate the influence of outliers in the data set. Furthermore, I extend the research of Bauer and Hamilton (2018) by considering the problems related to overlapping observations in the simulation study. I contribute to the existing literature by introducing new machine-learning methods which potentially have wide applications. Additionally, I resolve the spanning controversy by providing convincing evidence that there are additional predictors beyond the three yield PCs that have incremental predictive power for bond returns.

The remainder of this paper is structured as follows. Section 2 describes the data and sample periods. Section 3 contains a detailed description of the methodology. The main results are presented and analysed in Section 5. Ultimately, conclusions are drawn in Section 6.

2 Data

The monthly data sets used to revisit the studies of Joslin et al. (2014), Cieslak and Povala (2015), Cochrane and Piazzesi (2005), and Ludvigson and Ng (2009) are retrieved from the website of Michael Bauer[1]. Henceforth these four papers are abbreviated as JPS, CPO, CP, and LN respectively. The data sets are constructed by Bauer and Hamilton (2018) and contain data on the variables listed in Table 1 over the sample period used in the original paper. Bauer and Hamilton (2018) added data that was released after the publication of the original papers, resulting in an extension of the original sample periods to December 2016. In the remainder of this section, the data of the LN application and its extensions are described, as the focus of this paper is on these applications. The description of the variables used in JPS, CPO and CP is provided in Appendix A.

Table 1
The four papers that are revisited together with their variables and sample periods.

Paper	Variables	Original Sample
Joslin et al. (2014)	Economic growth, inflation and bond yields	1985-2008
Cieslak and Povala (2015)	Price level, one-month T-bill rate and bond yields	1971-2011
Cochrane and Piazzesi (2005)	Bond prices	1964-2003
Ludvigson and Ng (2009)	Bond prices and set of macroeconomic variables	1964-2007

From Table 1 it can be seen that Ludvigson and Ng (2009) use data on the bond prices of pure discount U.S Treasury bonds with maturities ranging from one to five years. These data are obtained from the Fama-Bliss data set from the Center for Research in Securities Prices (CRSP). Furthermore, they use a large set of 132 macroeconomic variables of which the data is provided by James Stock and Mark Watson. To ensure stationarity of the data, they apply transformations to the raw data. Thereafter, they standardize the transformed data. In order to extend the original sample period, Bauer and Hamilton (2018) use the data set from the website Michael McCracken[2]. The data sets provided on this website mimic the coverage of the Stock-Watson data sets and thus the information content in this data set is comparable to that of the data set employed by Ludvigson and Ng (2009).

Following Bauer and Hamilton (2018), I focus on both the original sample periods listed in Table 1 and the 1985-2016 sample in the empirical analysis for the four published papers. For LN and the new applications, I also consider the sample period starting from January 1985 until December 2022. To that end, an extended version of the large macroeconomic data set including monthly data up to December 2022 is retrieved from the website of Michael McCracken. In addition, the Fama-Bliss data set containing the monthly bond prices up to December 2022 is obtained from the CRSP. The reason to conduct part of the empirical analysis based on post-

[1]https://www.michaeldbauer.com/publication/robust-bond-risk-premia/.
[2]https://research.stlouisfed.org/econ/mccracken/fred-databases/.

1984 samples is that according to some papers the predictive power of macroeconomic variables for excess bond returns is weaker in more recent sample periods, especially in post-1984 samples. Additionally, monetary policy substantially changed in the early 1980s.

Table B.1 in Appendix B provides the list of the 128 variables included in the extended data set together with their descriptions. Despite the data set includes slightly fewer variables than that are considered by Ludvigson and Ng (2009), the large majority of the considered variables are the same. Similar to Ludvigson and Ng (2009), the data is transformed and standardized. The data transformations applied to each variable are also given in Table B.1. Furthermore, the data are cleaned in the same way as is done by Bauer and Hamilton (2018). Specifically, all variables with at least one missing observation in a certain sample period are excluded from the data set used for the empirical analysis over that sample period. Due to missing observations for the final months of 2016, some variables are removed from the data set employed by Bauer and Hamilton (2018) in their empirical analysis for the sample period 1986-2016. To avoid the duplication of these variables with potential predictive power for future excess bond returns, I also use the extended data set to study the performance of LN over this sample period.

The variables in the balanced data set are categorized into the eight groups defined by Ludvigson and Ng (2016): (1) output; (2) employment; (3) housing; (4) orders and inventories; (5) money market; (6) bond and foreign exchange (FX) market; (7) price indices; and (8) stock market. Table B.1 specifies to which of the eight groups each variable belongs. Even though I refer to this data set as the large macroeconomic data set, it is important to notice that the data set also contains financial variables, such as the variables categorized as stock market variables. Ludvigson and Ng (2009) argue that these variables should be included in the data set because fluctuations in the business cycle substantially co-move with financial and real macroeconomic variables. The common movements are presumed to be driven by shocks affecting both the aggregate economy and the financial markets. According to Ludvigson and Ng (2009), these joint variations are probable to be one of the most prominent sources of fluctuations in cyclical economic variables, such as bond risk premia, and therefore financial variables potentially contain predictive power for excess bond returns.

J. Z. Huang and Shi (2023) argue that some macroeconomic variables have a delayed impact on bond risk premia. For example, changes in the labour market and consumer prices appear to require a long lag before their effect becomes evident in the bond market. Therefore, following J. Z. Huang and Shi (2023), I also consider the data set that contains both the macroeconomic variables and their six lagged values. However, unless explicitly stated otherwise, the macroeconomic data set without lagged values is used in the empirical analysis.

3 Methodology

3.1 Predictive regression

To test the spanning hypothesis and to forecast bond returns, I consider the following predictive regression:

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

where y_{t+h} is the average excess return from buying certain bonds in month t and holding it h months, x_{1t} is a vector containing a constant and the values of the first three PCs of bond yields in month t, x_{2t} is a vector containing the values of certain additional predictors in month t, and u_{t+h} is the forecast error. The exact specifications for y_{t+h}, x_{1t} and x_{2t} used to revisit and extend the LN application are given in the next two sections, while those to revisit JPS, CPO and CP are given in Appendix C. The specific version of the spanning hypothesis that is tested in this paper is given by

$$H_0 : \beta_2 = 0. \tag{2}$$

If this null hypothesis is significantly rejected this would imply that there is statistical evidence that the variables in x_{2t} have predictive power for bond returns beyond the first three yield PCs.

3.2 Yield PCs and Average Excess Bond Returns

The price of a zero-coupon bond at time t that pays C_n in month $t+n$ is given by

$$P_t^{(n)} = C_n e^{-i_t^{(n)} \cdot n} \Leftrightarrow i_t^{(n)} = -\frac{1}{n} \log(P_t^{(n)}/C_n) \tag{3}$$

where $i_t^{(n)}$ is the continuously compounded monthly yield on the bond. Defining the log price of the n-month zero-coupon bond in period t as $p_t^{(n)} = \log(P_t^{(n)}/C_n)$, it follows that $i_t^{(n)} = -\frac{1}{n} p_t^{(n)}$. The three yield PCs in month t which are contained in x_{1t} are now obtained as

$$(PC_t^{(1)}, PC_t^{(2)}, PC_t^{(3)}) = W_{J\times 3}' i_t \tag{4}$$

where $i_t = (i_t^{(n_1)}, \ldots, i_t^{(n_J)})'$ are the bond yields from which the PCs are extracted and $W_{J\times 3}$ is the matrix that contains as columns the first three normalized eigenvectors corresponding to the variance matrix of i_t. In the LN application and its extensions, it holds that $i_t = (i_t^{(12)}, i_t^{(24)}, \ldots, i_t^{(60)})'$. These bond yields are computed from the data on the bond prices of the zero-coupon Treasury bonds with maturities from one to five years using Equation (3).

Using the above notation, the excess log return in period $t+h$ from buying an n-month

zero-coupon bond in period t and holding it h months is given by (Bauer & Hamilton, 2018)

$$rx_{t+h}^{(n)} = r_{t+h}^{(n)} - hi_t^{(h)} = (p_{t+h}^{(n-h)} - p_t^{(n)}) - hi_t^{(h)} = -(n-h)i_{t+h}^{(n-h)} + ni_t^{(n)} - hi_t^{(h)} \tag{5}$$

where $r_{t+h}^{(n)}$ is the log return from buying the n-month zero-coupon bond in period t and selling it as an $n-h$ month bond in period $t+h$. Now, the unweighted average of annual excess bond returns across bonds with maturities ranging from 2 to k years can be defined as

$$\overline{rx}_{t+12}^{(k)} = \frac{1}{k-1}\sum_{n=2}^{k} rx_{t+12}^{(12n)} \tag{6}$$

where $k \geq 2$. In the LN application and its extensions, the dependent variable in the predictive regression is the annual average excess bond return across bonds with maturities from two to five years, such that in these applications it holds that $y_{t+12} = \overline{rx}_{t+12}^{(5)}$.

However, using monthly data to estimate the predictive regression with annual returns as the dependent variable ($h = 12$) results in an econometric problem. Concretely, the overlapping returns result in $E(u_t u_{t+p} \neq 0)$ for $p = 0, 1, ..., 11$, and thus an MA(11) structure is induced for the forecast errors. According to Bauer and Hamilton (2018) this has two important implications. Firstly, in combination with persistent predictors, it substantially increases the variance of the ordinary least-squares (OLS) estimate for β_2 across different samples. Secondly, it substantially diminishes the reliability of the goodness of fit measure R^2, since including x_{2t} in the predictive regression can considerably increase the R^2 even if it has no predictive power. In the simulation study in Section 4, I illustrate these two implications.

As a standard attempt to correct for the correlation in the forecast errors, I use the Newey and West (1987) standard errors with 18 lags to calculate the test statistics. However, Ang and Bekaert (2006) showed that in case of overlapping returns the Newey-West standard errors are unreliable for testing the significance of regression coefficients and this inference is even less reliable as the persistence of the regressors increases. This is also demonstrated in the simulation study. As an alternative, the reverse-regression approach of Wei and Wright (2011) is used to calculate the standard errors in the CPO application. This approach uses the insight of Hodrick (1992) that regressing non-overlapping one-period returns on the sum of the predictors over the preceding h periods instead of regressing the h-period returns on the values of the predictors at the beginning of the holding period mitigates the problems arising from overlapping returns.

3.3 Factor Extraction Methods and their Predictor Factors

The large set of macroeconomic variables potentially contains variables with predictive power for excess bond returns beyond the first three yield PCs. However, including all the variables as additional predictors in the predictive regression would typically result in in-sample overfitting and poor out-of-sample performance. To reduce the dimension of the data set and avoid the curse of dimensionality, I use some well-known and self-constructed methods to extract one or multiple factors from the large data set. In the remainder of this section, the considered methods, to which I refer as factor extraction methods (FEM), are described. The input matrix for all these methods contains the standardized variables of the large macroeconomic data set over time and is denoted by

$$Z = \begin{pmatrix} z_1^{(1)} & z_1^{(2)} & \dots & z_1^{(N)} \\ z_2^{(1)} & z_2^{(2)} & \dots & z_2^{(N)} \\ \vdots & \vdots & \ddots & \vdots \\ z_T^{(1)} & z_T^{(2)} & \dots & z_T^{(N)} \end{pmatrix}$$

where T and N denote respectively the number of monthly observations and the number of variables in the data set.

3.3.1 PCA and sPCA Method

Following Ludvigson and Ng (2009), the first r PCs of the large macroeconomic data set are computed by multiplying $\sqrt{T}$ by the r eigenvectors corresponding to the largest r eigenvalues arranged in decreasing order of the sample covariance matrix $V_{zz}^{PCA} = \frac{1}{N}\sum_{i=1}^{N} z^{(i)}(z^{(i)})'$ with $z^{(i)} = (z_1^{(i)}, ..., z_T^{(i)})'$. These r eigenvectors are collected in the matrix $E_{T\times r}$ and are computed using singular value decomposition under the restriction that $E'_{T\times r}E_{T\times r} = I_r$ where I_r denotes the $r \times r$ identity matrix. To revisit LN, I follow Bauer and Hamilton (2018) and use the first eight macro PCs obtained as

$$F = (f_1, ..., f_8) = \sqrt{T}E_{T\times 8}$$

as additional predictors in the predictive regression. In addition, based on D. Huang et al. (2023), I consider the first six PCs as additional predictor factors and I use the fitted values of the regression of the target variable $\overline{rx}_{t+12}^{(5)}$ on the six PCs as an additional single predictor factor. I refer to these predictor factors as the **PCA** factors.

Even though PCA is helpful in reducing the dimensionality of a large data set, it completely disregards the target variable and therefore is not focussed on selecting factors that are most valuable for forecasting this variable (D. Huang et al., 2022). Since this research is about testing whether there are additional predictors beyond the first three yield PCs that have predictive

power for excess bond returns, it is interesting to consider factor selection methods that take the predictive power of the variables into account. To that end, D. Huang et al. (2022) propose the **sPCA** method which can be decomposed in two steps. In the first step, each variable i in the data set is scaled by $\hat{\gamma}_i$, where $\hat{\gamma}_i$ is the estimated coefficient in the regression of the target variable $\overline{rx}^{(5)}_{t+12}$ on variable i. As such, more weight is put on the variables with stronger predictive power for the target variable. In the second step, PCA is applied to the scaled data set $(\hat{\gamma}_1 z^{(1)}, \ldots, \hat{\gamma}_N z^{(N)})$. That is, the first r principal components of the matrix $V_{zz}^{sPCA} = \frac{1}{N}\sum_{i=1}^{N} \hat{\gamma}_i z^{(i)} (\hat{\gamma}_i z^{(i)})'$ are computed in the same way as in the procedure for the PCA method described above. Again, either the first six extracted factors or the fitted values of the regression of the target variable $\overline{rx}^{(5)}_{t+12}$ on the six factors are used as additional predictors x_{2t} in the predictive regression.

3.3.2 Machine-Learning Methods

Based on the intuition of the sPCA method and the method used in J. Z. Huang and Shi (2023) to obtain the SAGLasso factor, I construct several machine-learning methods to extract factors from the large macroeconomic data set. In total, I consider 12 methods of which some are existing methods and some are novel methods. The first method employs Lasso (Tibshirani, 1996) to extract a single predictor factor. In the other methods refined versions of this popular regression shrinkage method are used to obtain potential predictors. In the remainder of this section, Lasso and the considered refinements are discussed. The pseudocode in Appendix E gives an overview of all the considered machine-learning methods and describes how these methods are used to extract predictor factors from the large macroeconomic data set.

Lasso is a method that shrinks the coefficient estimates of a multivariate regression model towards zero by regularizing these estimates. Due to the nature of its regularization, it tends to set some coefficients exactly to zero (Tibshirani, 1996). As such, Lasso typically leads to parsimonious models and can be used to reduce the dimensionality of the large macroeconomic data set. Similar to the sPCA method it takes the target variable into account when selecting variables. To obtain the Lasso estimate of the coefficients of the regression with $\overline{rx}^{(5)}_{t+12}$ as dependent variable and the macroeconomic variables as independent variables the following minimization problem is solved:

$$\min_{\beta} \left(\sum_{t=1}^{T-12} \left(\overline{rx}^{(5)}_{t+12} - \sum_{i=1}^{N} z_t^{(i)} \beta_i \right)^2 + \sum_{i=1}^{N} \lambda |\beta_i| \right) \tag{7}$$

where λ is the tuning parameter of the penalty term, which is determined using five-fold cross-

validation (J. Z. Huang & Shi, 2023). The Lasso and OLS estimates if λ is zero.

The predictive information contained in the macroeconomic variables selected by Lasso possibly overlaps with the information in the yield curve. To minimize this information overlap, J. Z. Huang and Shi (2023) include the first three yield PCs in their variable selection method, but do not penalize the corresponding coefficients. Similarly, I include the three yield PCs in the first term of minimization problem (7), but do not include their coefficients in the penalty term of this minimization problem. The adjusted minimization problem corresponding to this first refinement of Lasso, referred to as Controlled Lasso (**CLasso**), is specified by

$$\min_{\beta} \left(\sum_{t=1}^{T-12} \left(\overline{rx}_{t+12}^{(5)} - \sum_{k=1}^{3} PC_t^{(k)} \beta_{1k} - \sum_{i=1}^{N} z_t^{(i)} \beta_{2i} \right)^2 + \sum_{i=1}^{N} \lambda |\beta_{2i}| \right) \tag{8}$$

where $PC_t^{(k)}$ is the value of kth yield PC in month t.

Zou (2006) points out that Lasso coefficient estimates can be biased. To solve this disadvantage of Lasso, he proposes to use Adaptive Lasso (**ALasso**) in which λ is replaced with λ_i in the minimization problem corresponding to Lasso and as a result the coefficients are penalized separately. Zou (2006) recommends to use OLS to determine λ_i. To avoid potential multicollinearity issues, I follow J. Z. Huang and Shi (2023) and use a ridge regression rather than OLS to obtain λ_i. In particular, I use the following specification for λ_i:

$$\lambda_i = \frac{\lambda}{|\hat{\beta}_i^{ridge}|^{\gamma}} \tag{9}$$

where $\hat{\beta}_i^{rigde}$ is the coefficient estimate for variable i of the ridge regression (Hoerl & Kennard, 1970) of $\overline{rx}_{t+12}^{(5)}$ on the macroeconomic variables. In J. Z. Huang and Shi (2023), γ and λ are jointly determined using cross-validation. I deviate from this approach by setting γ equal to one and using five-fold cross-validation only to compute λ. My approach is more efficient but may be less optimal.

Inspired by the sPCA method, I also apply PCA to the macroeconomic variables scaled by their (refined) Lasso coefficient estimates and use the resulting PCs as additional predictor factors x_{2t}. In case Lasso is applied in the first step, this method is referred to as Factor Lasso (**FLasso**) and is implemented as follows. Firstly, the coefficients of the regression of $\overline{rx}_{t+12}^{(5)}$ on the macroeconomic variables are estimated using Lasso yielding $\hat{\beta}$. Secondly, PCA as described in Section 3.3.1 is applied to the scaled macroeconomic data set $(\hat{\beta}_1 z^{(1)}, \ldots, \hat{\beta}_N z^{(N)})$. Furthermore, based on the successful SAGLasso algorithm employed by J. Z. Huang and Shi (2023), I construct two comparable yet distinctive two-step methods to extract predictor factors.

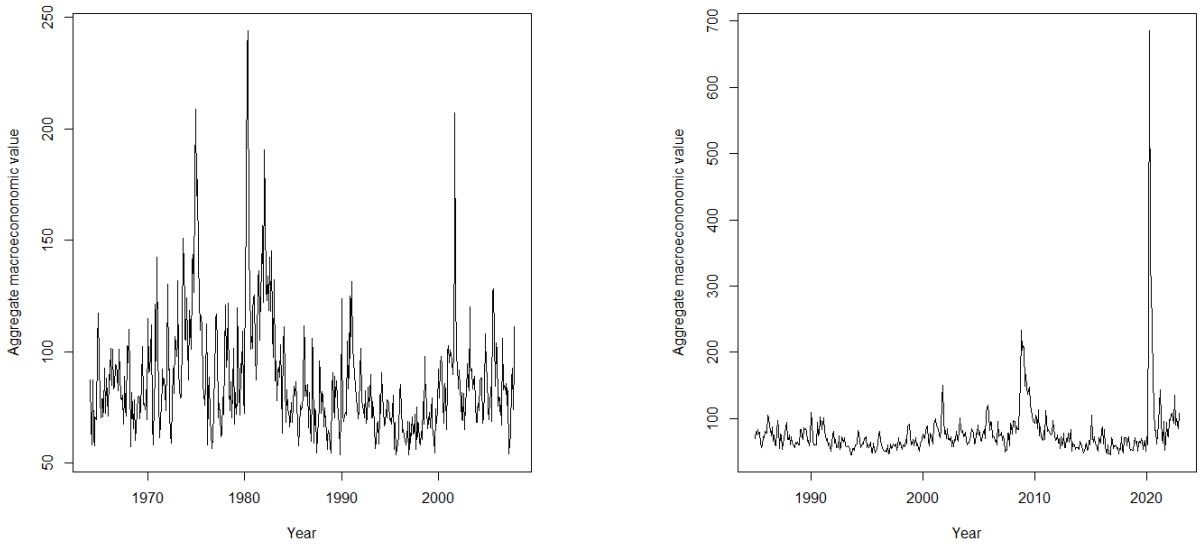
These methods, referred to as **GCALasso1** and **GCALasso2**, are described in Appendix D.

In the empirical analysis, I consider the following 12 machine-learning methods: Lasso, ALasso, CLasso, FLasso, CALasso, FALasso, FCLasso, FCALasso, GCALasso1, GCALasso2, FGCALasso1, and FGCALasso1. In each method, one or multiple approaches described above are combined and used to construct predictor factors x_{2t} from the large macroeconomic data set. As specified in the pseudocode in Appendix E, the Lasso, ALasso, CLasso, CALasso, GCALasso1 and GCALasso2 methods only produce a single predictor factor. The other machine-learning methods employ PCA at the end of their procedure and thus also produce either six individual predictor factors or a single predictor factor.

3.4 Winsorization Procedures

Figure 1 plots the sum of absolute values of the standardized macroeconomic variables in the large macroeconomic data set over the 1964-2007 sample and the 1985-2022 sample. From this figure, I identify several sudden large spikes. The peaks in the mid-1970s and early 1980s can be related to the recessions resulting from the oil shocks in 1973 and 1979. The extreme values in the last months of 2001 are caused by the September 11 attacks. The aberrant observations in 2008 and 2009 occurred during the Great Recession. Finally, the large spike at the beginning of 2020 corresponds to the start of the COVID-19 pandemic. This peak is so large that it dwarfs the other spikes.

To reduce the impact of the extreme values in the data set, outlier winsorizing techniques can be used. In contrast to traditional approaches, such techniques do not simply exclude outliers but mitigate the effect of outliers by adjusting their magnitude. Winsorizing techniques are especially valuable in scenarios where extreme values have a disproportionate effect on the results. In the empirical analysis, I consider two outlier winsorizing procedures. In both procedures 90% winsorization is applied, meaning that the 5% smallest values and the 5% largest values of a time series are replaced by the 5th and 95th percentile respectively. The first approach is based on Bottmer, Croux and Wilms (2022) and pre-processes the data by winsorizing the outliers of both the target variable and the macroeconomic variables before applying the factor extraction methods to the data. In the second procedure, the factor extraction methods are applied to the original data and the resulting predictor factors are winsorized afterwards. In the remainder of this paper, I refer to the first approach as input winsorization and to the second one as output winsorization. If one of these two winsorization methods is used in the empirical analysis, it is explicitly stated. However, in the default case no winsorization is applied.



(a) 1964-2007 **(b)** 1985-2022

Figure 1
The sum of the absolute values of the standardized macroeconomic variables contained in the extended large macroeconomic data set over the period from January 1964 to December 2007 (on the left) and the period from January 1985 to December 2022 (on the right).

3.5 Econometric problems

In addition to the econometric problem arising from the overlapping returns, Bauer and Hamilton (2018) argue that the predictive regression has a number of other econometric problems. Firstly, because the yield curve PCs in x_{1t} contain almost all information in the current yield curve, x_{1t} is necessarily correlated with u_t and thus x_{1t} is not strictly exogenous ($E(x'_{1t}u_t) = 0$). Furthermore, the first-order autocorrelation of the predictors is typically close to one, meaning that they are typically highly persistent. Bauer and Hamilton (2018) show that in small samples this leads to a downward bias in the standard errors of conventional tests and thus to size distortions. Moreover, they show that these size distortions are even larger when the considered predictors are trending over the sample. As a result, the spanning hypothesis is rejected too often. I further investigate the sources and effects of the standard error bias in the simulation study in Section 4.

To correctly determine the standard errors of β_2 in the predictive regression and to robustly test the spanning hypothesis, I use the bootstrap test and the corresponding code of Bauer and Hamilton (2018). In addition, bootstrapping under both the null hypothesis and the alternative hypothesis is used to assess the robustness of the conventional and bootstrap tests. Finally, the bias correction proposed by Kilian (1998) is implemented in the bootstrap procedure for the JPS and CPO applications, because x_{2t} is very persistent in these applications leading to a bias in the simple bootstrap. I refer to this adjusted bootstrap as the "bias-corrected bootstrap". In Appendix F the bootstrap procedures are described in detail.

3.6 Out-of-sample forecasting

In addition to performing in-sample tests of the spanning hypothesis, the out-of-sample forecasting performance of each considered model is evaluated. Following most related literature, expanding window estimation is used in the out-of-sample analysis. Bauer and Hamilton (2018) also estimate the parameters of the predictive regression model recursively using expanding windows, but they do not recursively estimate the yield PCs and the PCs extracted from a large macroeconomic data set. Instead, they estimate these PCs only once over the full sample, including the out-of-sample period, and use these PCs to construct out-of-sample forecasts. As a result, the out-of-sample results obtained by (Bauer & Hamilton, 2018) suffer from a look-ahead bias. To correct for this bias, I estimate the yield PCs and the additional predictor factors as well as the parameters recursively using data only through month t to calculate the 12-month ahead forecast for month $t + 12$.

For the four studies to be revisited, the corresponding original sample as listed in Table 1 is used as the initial estimation window. In the case of the other applications, the initial estimation window starts in January 1964 and ends in December 2007. The mean-squared-errors of the resulting forecasts from both the restricted model under the null hypothesis ($\beta_2 = 0$) and the unrestricted model ($\beta_2 \neq 0$) are compared by means of the modified Diebold and Mariano (2002) test proposed by Harvey, Leybourne and Newbold (1997). This modified test accounts for the presence of overlapping observations and the resulting autocorrelation in the prediction errors.

4 Simulation study

To understand the sources and effects of the econometric problems underlying the testing of the spanning hypothesis in small samples, I run several Monte Carlo simulation experiments. Firstly, I replicate the simulation exercise of Bauer and Hamilton (2018) which examines the effects of the presence of endogeneity, persistent regressors and trending regressors in the predictive regression, but ignores the presence of overlapping observations. The procedure and results for this simulation exercise are described in Appendix G.1. From the analysis, it can be concluded that the presence of endogeneity, persistent regressors and trending regressors in the predictive regression leads to a standard error bias in the conventional tests. The standard error bias in $\hat{\beta}_2$ causes size distortions which increase in the persistence of the regressors, the correlation between x_{1t} and u_t, and the trend in x_{2t}. Furthermore, it follows that the bootstrap test is relatively robust to these small-sample econometric problems and does not lead to large size distortions. The details of this analysis are given in the Appendix.

In addition, I extend the simulation study of Bauer and Hamilton (2018) by imposing a MA(11) structure on the error terms of the dependent variable of the predictive regression. As a result, the simulation study matches more closely the empirical setting with overlapping observations that is encountered in Bauer and Hamilton (2018) and other studies investigating the spanning hypothesis. The procedure and results of this simulation exercise are described in the next subsection.

4.1 Overlapping observations

To quantify the magnitude of the effect of overlapping observations, I use the following DGP to generate simulation data:

$$x_{it} = \mu_i + \rho_i x_{i,t-1} + \epsilon_{it} \qquad i = 1, 2 \qquad t = 1, ..., T + 12 \tag{10}$$

$$y_t = \epsilon_t^y = v_{3t} + \frac{1}{2} \sum_{s=t-h+1}^{t-1} v_{3s} \qquad t = 12, ..., T + 12 \tag{11}$$

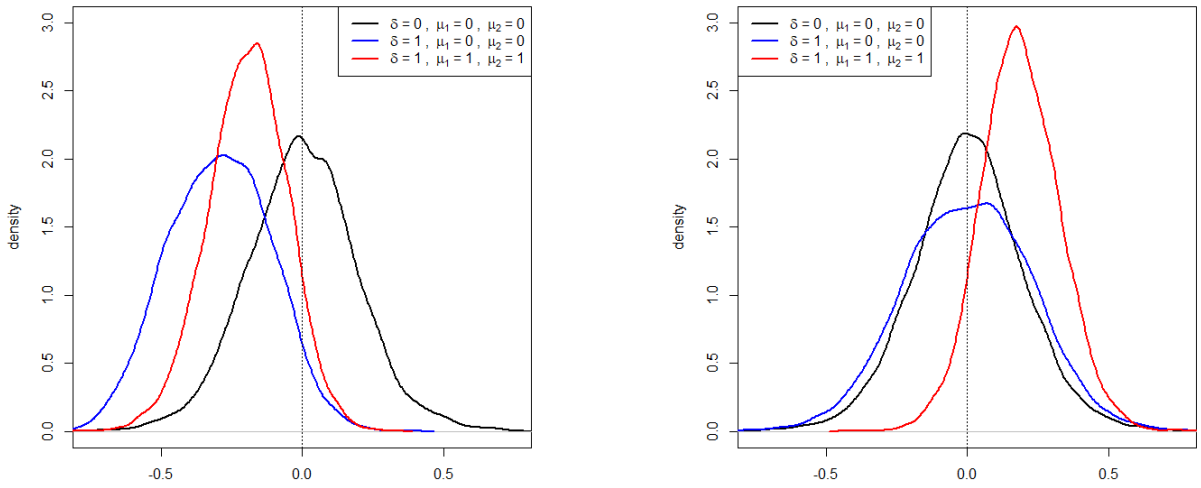
where $x_{10} = x_{20} = 0$, $\epsilon_{1t} = \delta v_{3t} + \sqrt{1 - \delta^2} v_{1t}$, $\epsilon_{2t} = v_{2t}$, and $v_{it} \overset{iid}{\sim} N(0, 1)$ for $i \in \{1, 2, 3\}$. Consequently, an MA(11) structure is imposed on the error terms ϵ_t^y. $N_s = 10,000$ samples of length $T = 100$ are generated and used to examine the small sample properties of the predictive regression $y_{t+12} = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + u_{t+12}$ with $t = 1, ..., 100$. In this setting, Newey-West standard errors with 18 lags are used instead of OLS standard errors. In each simulation experiment, the correlation between x_{1t} and ϵ_{t+12}^y, the coefficient and standard error bias of $\hat{\beta}_1$ and $\hat{\beta}_2$, the standard deviation of $\hat{\beta}_2$, the size of the test of $H_0 : \beta_2 = 0$ for the conventional t-test and bootstrap test, and the average and standard deviation of the difference in R^2 of the predictive regressions under the alternative hypothesis (R_2^2) and the null hypothesis (R_1^2) are computed. The formulas and procedures used to compute these statistics are given in Section G.2.

The results for this simulation exercise are shown in Table 2. Comparing these results to the results obtained in the simulation exercise without overlapping observations leads to the following conclusions. In general, in the presence of overlapping observations, the patterns described in detail in Appendix G.1 are the same. In particular, the size distortions in the presence of overlapping observations also increase in ρ_i, δ and μ_2. However, in this case, the negative correlations between x_{1t} and the error term of the simulated dependent variable are larger leading to larger biases in in the standard errors of $\hat{\beta}_2$. For that reason, the size distortions of the conventional test are even larger and it can be concluded that the Newey-West standard errors are not able to overcome the problems emerging from overlapping returns in small samples.

Table 2
Simulation results for the basic setting without overlapping returns. In this simulation, 10,000 simulation samples of length $T = 100$ are generated according to the data-generating process (DGP) specified in Equations (10) and (11) for different values of δ, ρ_i and μ_i, $i = 1, 2$. In each simulation sample, the predictive regression $y_{t+12} = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + u_{t+12}$ is estimated. The statistics corresponding to the 10,000 regression results are reported in the table. The table reports the correlation between x_{1t} and ϵ^y_{t+12}, the coefficient and standard error bias of $\hat{\beta}_1$ and $\hat{\beta}_2$, the standard deviation of $\hat{\beta}_2$, the size of the test of $H_0 : \beta_2 = 0$ for the conventional t-test and bootstrap test, and the average and standard deviation of the difference in R^2 of the predictive regressions under the alternative hypothesis (R^2_2) and the null hypothesis (R^2_1) are computed. The formulas and procedures used to compute these statistics are given in Appendix G.2.

			Corr.	Coefficient Bias		SE bias (%)		Std.	Size		$R^2_2 - R^2_1$	
ρ_1	ρ_2	δ	$(x_{1t}, \epsilon^y_{t+12})$	β_1	β_2	β_1	β_2	β_2	Simulated	Bootstrap	Mean	Std.
$\mu_1 = \mu_2 = 0$												
0.99	0.99	0.0	0.002	0.001	0.004	-43.9	-43.8	0.196	0.298	0.054	0.07	0.09
0.00	0.00	1.0	-0.044	-0.059	0.001	-21.5	-19.0	0.187	0.118	0.048	0.01	0.01
0.99	0.00	1.0	-0.308	-0.201	0.003	-43.2	-20.9	0.173	0.126	0.051	0.01	0.01
0.99	0.80	1.0	-0.310	-0.205	-0.000	-45.5	-37.1	0.257	0.242	0.059	0.04	0.06
0.90	0.90	1.0	-0.188	-0.202	-0.001	-39.5	-42.5	0.265	0.298	0.071	0.07	0.09
0.99	0.99	0.8	-0.311	-0.235	0.000	-46.7	-50.3	0.214	0.369	0.076	0.09	0.11
0.99	0.99	1.0	-0.388	-0.292	0.003	-49.8	-54.4	0.225	0.424	0.107	0.10	0.12
$\mu_1 = 0$, $\mu_2 = 1$												
0.99	0.99	0.0	-0.000	0.002	0.000	-44.5	-48.0	0.042	0.326	0.053	0.09	0.11
0.00	0.00	1.0	-0.044	-0.059	0.001	-21.9	-18.7	0.188	0.117	0.051	0.01	0.01
0.99	0.00	1.0	-0.312	-0.203	0.002	-43.3	-20.1	0.171	0.120	0.051	0.01	0.01
0.99	0.80	1.0	-0.307	-0.205	0.002	-45.7	-36.6	0.243	0.236	0.054	0.04	0.06
0.90	0.90	1.0	-0.187	-0.210	0.001	-40.2	-45.1	0.206	0.326	0.064	0.08	0.09
0.99	0.99	0.8	-0.308	-0.305	0.001	-43.7	-62.4	0.056	0.478	0.100	0.14	0.14
0.99	0.99	1.0	-0.388	-0.379	0.000	-44.3	-67.3	0.063	0.571	0.134	0.18	0.16
$\mu_1 = 1$, $\mu_2 = 0$												
0.99	0.99	0.0	-0.003	-0.001	0.001	-46.7	-44.6	0.215	0.306	0.058	0.07	0.09
0.00	0.00	1.0	-0.045	-0.062	-0.000	-21.6	-18.1	0.186	0.117	0.048	0.01	0.01
0.99	0.00	1.0	-0.053	-0.006	-0.001	-42.0	-19.9	0.178	0.125	0.051	0.01	0.01
0.99	0.80	1.0	-0.055	-0.006	-0.006	-45.0	-36.7	0.259	0.236	0.054	0.04	0.06
0.90	0.90	1.0	-0.142	-0.116	-0.000	-40.5	-41.5	0.255	0.285	0.063	0.07	0.08
0.99	0.99	0.8	-0.052	-0.009	0.000	-48.5	-44.5	0.214	0.307	0.056	0.07	0.09
0.99	0.99	1.0	-0.059	-0.011	0.003	-48.2	-45.8	0.219	0.314	0.054	0.07	0.09
$\mu_1 = 1$, $\mu_2 = 1$												
0.99	0.99	0.0	0.001	0.000	-0.000	-44.5	-44.5	0.153	0.300	0.046	0.07	0.09
0.00	0.00	1.0	-0.046	-0.062	0.001	-22.3	-19.1	0.189	0.123	0.054	0.01	0.01
0.99	0.00	1.0	-0.054	-0.006	-0.001	-42.1	-20.6	0.177	0.125	0.052	0.01	0.01
0.99	0.80	1.0	-0.054	-0.006	0.019	-45.4	-36.2	0.249	0.237	0.056	0.04	0.06
0.90	0.90	1.0	-0.145	-0.149	0.065	-41.8	-42.6	0.211	0.315	0.069	0.07	0.09
0.99	0.99	0.8	-0.055	-0.154	0.149	-44.7	-44.6	0.143	0.505	0.102	0.13	0.13
0.99	0.99	1.0	-0.063	-0.193	0.187	-46.2	-45.7	0.139	0.620	0.138	0.16	0.14

Additionally, the sampling variability of $\hat{\beta}_2$ across the simulated samples is substantially increased in this case. For example, the standard deviation of the OLS estimates for $\hat{\beta}_2$ is about four times larger if $\mu_i = 0$, $\rho_i = 0$, and $\delta = 1$. This can also be seen when comparing Figure 2 with Figure G.1. The figures are similar but the scale of the axes differ and it follows that the distributions of $\hat{\beta}_1$ and $\hat{\beta}_2$ are wider in the presence of overlapping observations. Finally, both the average and standard deviation of $R^2_2 - R^2_1$ are increased, meaning that the reliability of this measure is considerably reduced when the error terms in the regression are serially correlated. These conclusions are in line with the theoretical results derived in Bauer and Hamilton (2018). Nevertheless, the size distortions of the bootstrap test are also limited in the presence of overlapping observations. The bootstrap test is somewhat oversized if both regressors are persistent, but in this situation the bias-corrected bootstrap can be employed to alleviate the problem relating to the estimation of persistent processes.



(a) β_1 **(b)** β_2

Figure 2
Simulation distibution of $\hat{\beta}_1$ (on the left) and $\hat{\beta}_2$ (on the right) for three different scenarios. In this simulation, 10,000 simulation samples of length $T = 100$ are generated according to the data-generating process (DGP) specified in Equations (10) and (11) with $\rho_1 = \rho_2 = 0.99$ and for different values of δ, μ_1 and μ_2. In each simulation sample, the predictive regression $y_{t+12} = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + u_{t+12}$ is estimated and the resulting $\hat{\beta}_1$ and $\hat{\beta}_2$ are included in the density plots.

5 Empirical results

In this section, the in-and-of-of-sample predictive ability for excess bond returns beyond the three yield PCs of the proposed additional predictors is empirically evaluated. I focus on the LN application and the factor extraction methods that consistently perform best either in-sample or out-of-sample and present their results in this section. The results of the other factor extraction methods are reported in the Appendix. In this way, space can be conserved for more extensive and clear interpretations of the results. The in-sample analysis is described in Section 5.1. To revisit the evidence provided by Ludvigson and Ng (2009), I follow Bauer and Hamilton (2018) and use the first eight PCs of the large macroeconomic data set as the additional predictors in x_{2t}. For the other factor extraction methods, I follow D. Huang et al. (2023) and only consider the single predictor factor resulting from each method as the additional predictor x_{2t} in the in-sample analysis. The rationales behind the latter decision are provided when discussing the in-sample results corresponding to the LN application. In the out-of-sample analysis described in Section 5.2, I take into consideration both the single predictor factor and the individual factors resulting from the methods as additional predictors x_{2t} in the predictive regressions.

To assess the magnitude of the econometric problems in the above-mentioned applications as well as in the JPS, CPO and CP applications, some preliminary analysis is conducted in Appendix H. The analysis illustrates that the econometric problems are more severe in some applications than in others. Additionally, it substantiates the decision to employ the bias-

correction in the bootstrap for JPS and CPO, while omitting it from the bootstrap for the other applications. Ultimately, it shows that the values of the parameters used in the simulation study provide a realistic depiction of those that may be encountered in real-world scenarios. For the details, I refer to the Appendix.

The in-and-out-of-sample empirical analyses for JPS, CPO and CP are conducted in Appendix I. In the examination of the evidence provided by these published studies, I closely follow Bauer and Hamilton (2018). From the in-sample analysis, it follows that the size distortions of the conventional tests are large, especially for JPS and CPO, indicating that these tests are unreliable for inference in the predictive regressions. The bootstrap test proposed by Bauer and Hamilton (2018) is robust to the small-sample econometric problems and using this test to perform inference leads to weaker evidence against the spanning hypothesis than is suggested in the revisited papers and often even leads to insignificant results. Considering recent data even further weakens the evidence. The predictive power for excess bond returns of the additional predictors x_{2t} proposed by these three studies is thus far from convincing. The out-of-sample results corroborate this conclusion. In particular, adding the additional predictors x_{2t} to a forecasting model with only the three yield PCs leads to significantly higher prediction errors.

5.1 In-sample Analysis

5.1.1 LN application

The in-sample coefficient estimates and statistics of the eight macro PCs as additional predictors x_{2t} in the predictive regressions estimated over the samples 1964-2007 and 1985-2022 are given in the first panel of Table 3. These results are used to formally test the spanning hypothesis. According to the conventional p-values, five macro PCs are significant at 10% level and three of them are significant at 5% level in the original LN sample. The evidence is weaker when applying the simple bootstrap test. In this case, three PCs are significant at 10% level and only one factor at 5% level. The joint significance of the eight macro PCs is tested using the Wald test. The Wald p-values corresponding to both the conventional test and the bootstrap test indicate that the eight macro PCs jointly are highly significant in the original LN sample, but again the evidence is somewhat weaker when using the bootstrap test. The bootstrap test thus results in less compelling evidence against the spanning hypothesis than is suggested by Ludvigson and Ng (2009), but the evidence is still statistically sufficient to support the claim that at least some macro PCs are helpful in predicting excess bond returns in this sample.

Table 4 reports the adjusted R^2 for the restricted regression model with only x_{1t} as predictors and the unrestricted regression model with both x_{1t} and x_{2t} as predictors in the actual data

Table 3
The in-sample coefficients and statistics of the additional predictors x_{2t} in the predictive regressions corresponding to LN and the other relevant factor extraction methods (FEM) estimated over the samples 1964-2007 and 1985-2022. In each regression, the dependent variable is $\overline{rx}^{(5)}_{t+12}$ and x_{1t} consists of a constant and the first three yield curve PCs. For LN, x_{2t} contains the first eight PCs extracted from a large macroeconomic data set. The eight macro PCs provided by Ludvigson and Ng (2009) are utilized in the regression for the original sample period, while the eight macro PCs extracted from the extended macroeconomic data set are used for the later sample period. In the case of the other factor extraction methods, x_{2t} is the single predictor factor extracted from the extended macroeconomic data set for both sample periods. No winsorization is applied when extracting the additional predictors. Under *Wald* the table reports the statistics corresponding to the test of joint significance of the eight macro PCs in the case of LN. The conventional statistics and p-values are computed using Newey-West standard errors with 18 lags. The simple bootstrap procedure is employed in all the applications. The conventional size and power are estimated using Equation (26) and (29) respectively. The bootstrap 5% critical values (c.v.'s) and p-values are computed using Equation (25) and (24) respectively. The size and power of the bootstrap test are approximated using Equations (27) and (30). The p-values that are lower than 5% are highlighted in bold.

		Stat.	5% c.v.	p-value		Size		Power	
	Coefficient	Conv.	Bootstrap	Conv.	Bootstrap	Conv.	Bootstrap	Conv.	Bootstrap
LN									
1964-2007									
f_1	0.742	1.855	2.458	0.064	0.136	0.115	0.048	0.625	0.459
f_2	0.147	0.380	2.498	0.704	0.746	0.120	0.048	0.143	0.063
f_3	0.072	0.608	2.246	0.544	0.584	0.092	0.050	0.125	0.084
f_4	-0.528	-1.912	2.544	0.056	0.135	0.128	0.053	0.590	0.420
f_5	-0.321	-1.307	2.493	0.192	0.289	0.117	0.048	0.328	0.203
f_6	0.576	2.221	2.645	**0.027**	0.094	0.137	0.048	0.538	0.366
f_7	0.401	2.361	2.493	**0.019**	0.061	0.118	0.055	0.637	0.516
f_8	0.551	3.036	2.322	**0.003**	**0.012**	0.099	0.049	0.849	0.791
Wald		42.073	29.077	**0.000**	**0.009**	0.322	0.054	0.998	0.962
1985-2022									
f_1	-0.448	-2.847	2.520	**0.005**	**0.028**	0.125	0.055	0.876	0.795
f_2	-0.501	-2.860	2.859	**0.004**	**0.050**	0.168	0.047	0.745	0.530
f_3	-1.145	-3.028	3.136	**0.003**	0.058	0.195	0.056	0.813	0.601
f_4	0.279	1.254	2.552	0.210	0.316	0.121	0.055	0.350	0.203
f_5	0.086	0.812	2.451	0.417	0.513	0.108	0.054	0.177	0.103
f_6	0.155	0.458	2.672	0.647	0.728	0.139	0.057	0.185	0.073
f_7	0.033	0.141	2.475	0.888	0.915	0.112	0.052	0.126	0.054
f_8	-0.042	-0.267	2.323	0.790	0.819	0.096	0.052	0.122	0.080
Wald		29.067	34.834	**0.000**	0.091	0.408	0.056	0.941	0.645
FEM									
1964-2007									
PCA	0.675	2.984	2.516	**0.003**	**0.023**	0.118	0.055	0.901	0.827
sPCA	0.705	3.130	2.700	**0.002**	**0.026**	0.142	0.049	0.891	0.774
FLasso	0.635	2.746	2.418	**0.006**	**0.028**	0.108	0.052	0.904	0.840
GCALasso1	4.072	10.679	2.444	**0.000**	**0.000**	0.108	0.049	1.000	1.000
GCALasso2	4.043	9.789	2.463	**0.000**	**0.000**	0.121	0.045	1.000	1.000
1985-2022									
PCA	1.171	2.818	2.652	**0.005**	**0.040**	0.132	0.052	0.913	0.807
sPCA	0.909	3.462	2.344	**0.001**	**0.005**	0.101	0.057	0.998	0.993
FLasso	0.790	3.407	2.362	**0.001**	**0.003**	0.096	0.051	0.999	0.995
GCALasso1	3.121	8.417	2.795	**0.000**	**0.000**	0.151	0.055	1.000	1.000
GCALasso2	3.114	7.781	2.867	**0.000**	**0.000**	0.163	0.056	1.000	1.000

sets as well as its mean and 95%-quantiles in the bootstrap samples generated under the null hypothesis that x_{2t} has no predictive power. Furthermore, the table reports the results for the increase in the adjusted R^2 when adding x_{2t} to the restricted model. The results for LN indicate that the adjusted R^2 increases from 0.25 to 0.35 in the original sample when including the eight macro PCs in the restricted forecasting model. Even though this increase of 10 percentage points seems quite large, it falls within the 95% bootstrap interval and thus is not significant at 5% level. The increase in the adjusted R^2 is thus not implausible under the null hypothesis and this casts some doubt on the additional predictive ability of the eight macro PCs in the original sample.

From Table 3 it also follows that in this application the conventional tests are oversized, meaning that their true size is above the nominal size of 5%. The size distortions of the conventional t-tests are relatively small. In the original sample, the true sizes of these t-tests range between 9.2% and 13.7%. The estimate of the true size of the conventional Wald test is substantially larger and is equal to 32.2% in the original sample. According to Bauer and Hamilton (2018) the larger size distortion in the case of the conventional Wald test is due to the fact that the Wald test compounds the econometric problems related to each of the eight individual conventional t-tests. On the contrary, the sizes corresponding to the bootstrap tests are all close to 5%. The results for the 1985-2022 sample are similar and lead to the same conclusions.

The table also reports the estimates for the power of the conventional and bootstrap tests. In the procedure to calculate these estimates, bootstrap samples are generated under the alternative hypothesis by adding $\hat{\beta}_2 \tilde{x}_{2\tau}$ to $\tilde{y}_{\tau+h}$, where $\hat{\beta}_2$ is the estimated coefficient of β_2 in the predictive regression using the actual data, $\tilde{x}_{2\tau}$ are the bootstrapped additional predictors and $\tilde{y}_{\tau+h}$ is the dependent variable bootstrapped under the null hypothesis. As such, the bootstrap samples are generated under the assumption that the additional predictors x_{2t} predict the target variable y_{t+h} with the magnitude that follows from the actual data. The power is now estimated as the fraction of bootstrap samples in which the null hypothesis is rejected. If this fraction is low, this does not mean that the test lacks power, but it is an indication that the additional predictors x_{2t} do not have significant predictive power for excess bond returns beyond the three yield PCs. The detailed procedure to calculate the power is described in Appendix F.2.

From the table, it can be seen that the estimates for the power of the conventional test are always larger than those of the bootstrap test. Even in case the coefficient estimates are statistically insignificant, the estimate for the power is relatively large for the conventional test. For example, the eight macro PCs jointly are not significant at 5% level according to the bootstrap test in the 1985-2022 sample, but adding them to the bootstrapped dependent variable still leads to a rejection of the null hypothesis in 94% of the bootstrap samples when using the conventional test compared to 65% when using the bootstrap test. This is another indication that the conventional test is less strict in terms of rejecting the spanning hypothesis and that evidence against the spanning hypothesis resulting from these tests should be interpreted with caution. Together with the results regarding the size of the tests and in line with Appendix I, it can be concluded that the conventional tests are unreliable for inference about bond risk premia. Conversely, the bootstrap procedure proposed by Bauer and Hamilton (2018) provides a robust alternative to test the spanning hypothesis.

Similar to the other applications discussed in Appendix I, the evidence against the spanning

Table 4
In-sample adjusted R^2 for the restricted regression model with only x_{1t} (R_1^2), the adjusted R^2 for the unrestricted regression model including both x_{1t} and x_{2t} (R_2^2), and the difference in adjusted R^2 ($R_2^2 - R_1^2$) corresponding to the LN application and the relevant other factor extraction methods estimated over the samples 1964-2007 and 1985-2022. In each regression model, the dependent variable is $\overline{rx}_{t+12}^{(5)}$ and x_{1t} consists of a constant and the first three yield curve PCs. For LN, x_{2t} contains the first eight PCs extracted from a large macroeconomic data set. The eight macro PCs provided by Ludvigson and Ng (2009) are utilized in the regression for the original sample period, while the eight macro PCs extracted from the extended macroeconomic data set are used for the later sample period. In the case of the other factor extraction methods, x_{2t} is the single predictor factor extracted from the extended macroeconomic data set for both sample periods. No winsorization is applied when extracting the additional predictors. The left half of the table provides the results for the earlier sample periods and the right half of the table provides the results for the later sample periods. For each application, the first row reports the adjusted R^2 statistic in the corresponding actual data set; the second and third rows report respectively the mean and 95%-quantiles of the statistics in the 5,000 bootstrap replications under H_0.

		R_1^2	R_2^2	$R_1^2 - R_2^2$	R_1^2	R_2^2	$R_1^2 - R_2^2$
		Earlier sample, 1964-2007			Later sample, 1985-2022		
LN	Data	0.25	0.35	0.10	0.16	0.26	0.10
	Bootstrap	0.21	0.24	0.03	0.32	0.35	0.04
		(0.05, 0.38)	(0.08, 0.42)	(-0.00, 0.11)	(0.12, 0.52)	(0.15, 0.55)	(-0.00, 0.12)
PCA	Data	0.25	0.32	0.07	0.16	0.22	0.06
	Bootstrap	0.20	0.21	0.01	0.32	0.33	0.01
		(0.05, 0.39)	(0.06, 0.40)	(-0.00, 0.05)	(0.12, 0.53)	(0.13, 0.53)	(-0.00, 0.05)
sPCA	Data	0.25	0.32	0.07	0.16	0.23	0.07
	Bootstrap	0.20	0.22	0.01	0.32	0.32	0.00
		(0.05, 0.39)	(0.06, 0.41)	(-0.00, 0.07)	(0.12, 0.52)	(0.12, 0.53)	(-0.00, 0.02)
FLasso	Data	0.25	0.29	0.04	0.16	0.23	0.07
	Bootstrap	0.20	0.21	0.01	0.32	0.32	0.00
		(0.05, 0.39)	(0.06, 0.40)	(-0.00, 0.04)	(0.12, 0.53)	(0.12, 0.53)	(-0.00, 0.02)
GCALasso1	Data	0.25	0.56	0.30	0.16	0.54	0.38
	Bootstrap	0.20	0.21	0.01	0.32	0.33	0.01
		(0.06, 0.39)	(0.06, 0.39)	(-0.00, 0.05)	(0.12, 0.52)	(0.13, 0.53)	(-0.00, 0.08)
GCALasso2	Data	0.25	0.57	0.31	0.16	0.50	0.34
	Bootstrap	0.20	0.21	0.01	0.32	0.33	0.02
		(0.05, 0.39)	(0.06, 0.40)	(-0.00, 0.05)	(0.12, 0.53)	(0.14, 0.54)	(-0.00, 0.08)

hypothesis is weaker in the post-1985 samples. The right part of Table 4 shows the results regarding the adjusted R^2 for the regressions in the 1985-2022 sample. The unreported results for the 1985-2016 sample are similar. For both samples, it holds that the increase in the adjusted R^2 resulting from adding the eight macro PCs as additional predictors x_{2t} to the restricted regression model is not significant at 5% level. The in-sample coefficient estimates and statistics of the eight macro PCs in the predictive regression estimated over the sample 1985-2016 are reported in Table J.1 in the Appendix. According to the bootstrap Wald test the eight macro PCs are far from being jointly significant. In addition, only one macro PC is significant at 10% level and none at 5% level. Compared to the 1985-2016 sample, the weakening of the evidence against the spanning hypothesis is less pronounced in the 1985-2022 sample. As is shown in Table 3, the bootstrap Wald p-value in this sample is equal to 9.1% such that the eight macro PCs are jointly significant at 10%. Furthermore, three PCs are individually significant at 10% level according to the bootstrap test in this sample.

Moreover, the individual macro PCs that are significant at 10% level according to the bootstrap test are different in the post-1985 samples. The last three macro PCs are significant in the 1964-2007 sample, while the first three macro PCs are significant in the 1985-2022 sample. There are two possible explanations for this. The first potential explanation is that the interpretations

of the last three macro PCs in the earlier sample are similar to the interpretations of the first three macro PCs in the later sample. This would suggest that the macroeconomic information that contains additional predictive power beyond the three yield PCs is stable over time, even though this predictive power is somewhat weaker in the later sample. The other potential explanation is that the macroeconomic information that contains additional predictive power for excess bond returns has changed over time. In any case, the joint and individual predictive power of the eight macro PCs is sample-specific and this corroborates the conclusion of Duffee (2013) that the results in Ludvigson and Ng (2009) lack stability across different samples.

One disadvantage of using PCs as predictors is their potential lack of economic interpretability. As such, it can be hard to link the predictive power of the macro PCs to specific macroeconomic concepts. On top of that, the interpretation of a certain PC may be different for distinct sample periods. Ludvigson and Ng (2009) argue that the predictive power of the first macro PC in their sample can be mapped to real activity, because the factor is highly correlated with measures of economic activity, such as industrial production. However, Duffee (2013) shows that only a small portion of the predictive power in the eight macro PCs can be attributed to measures of real activity and that the other macro PCs are harder to interpret.

To avoid the difficulties related to interpreting eight macro PCs separately and quantifying the predictive power of each macro PC in the different sample periods, I follow D. Huang et al. (2023) and focus in the in-sample analysis for the other factor extraction methods on the additional predictive power of the single predictor factor. Importantly, considering the single predictor factors instead of the multiple individual factors as additional predictors x_{2t} in the predictive regression simplifies interpretation without yielding worse out-of-sample performance, as in Section 5.2 it will be shown that the out-of-sample performance is generally very similar in both cases. An additional advantage of this approach is that the formal in-sample testing of the spanning hypothesis in this case boils down to testing the significance of only one coefficient in the predictive regression. As such, there is no need for a Wald test for the joint significance of multiple coefficients which has the drawback that it compounds the econometric problems associated with each of the individual coefficient estimates.

5.1.2 Other Relevant Factor Extraction Methods

In this section, I assess whether the single predictor factors resulting from the factor extraction methods described in Section 3.3 have in-sample predictive power for bond returns above the three yield curve PCs. In the analysis, I focus on results for the most interesting methods, namely the PCA, sPCA, FLasso, GCALasso1 and GCALasso2 methods. The second panel of

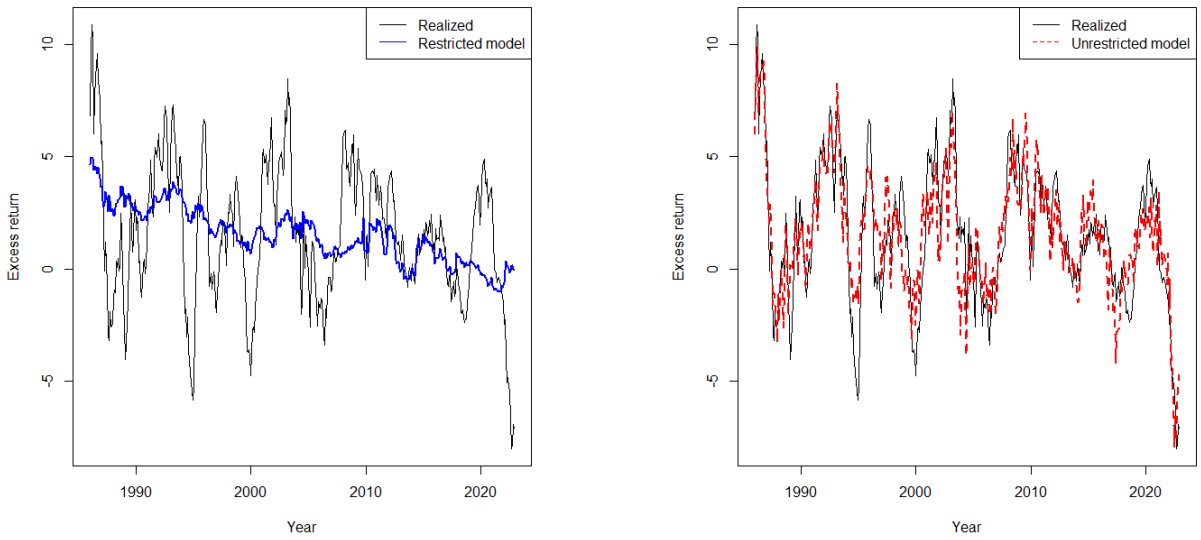
Table 3 reports the coefficient estimates and statistics of the single predictor factor resulting from each of the five methods as the additional predictor x_{2t} in the predictive regressions for the sample periods 1964-2007 and 1985-2022. It follows that the true sizes of the conventional test are larger than 5%, but the size distortions are modest compared to other applications. The true sizes of the bootstrap test are approximately equal to 5% and thus the bootstrap test is also a more robust test for the spanning hypothesis in these applications.

All the single predictor factors listed in the table are significant at 5% level in both the earlier and the later sample, even according to the bootstrap test. Especially the GCALasso1 and GCALasso2 factors are strongly significant with bootstrap p-values lower than 0.1%. Table 4 shows that the inclusion of the single predictor factor resulting from the PCA, GCALasso1 and GCALasso2 methods in the restricted forecasting model leads to significant increases in the adjusted R^2. The improvement in the adjusted R^2 due to adding either the GCALasso1 or GCALasso2 factor is about 30 percentage points in the earlier sample and 35 percentage points in the later sample. Since the improvements fall far outside the corresponding 95% bootstrap intervals, it is implausible that the increases in the adjusted R^2 are entirely spurious. Adding the single predictor factor of the sPCA and FLasso methods to the restricted forecasting model also leads to significant results in the later sample, but not in the earlier sample.

The above-mentioned results are obtained without the use of winsorization. I have also applied input and output winsorization in the process of extracting factors. Even though the results are slightly better for some methods, it generally leads to the same conclusions and therefore these results are untabulated for brevity. Overall, it can be concluded that winsorization does not help (much) to improve the in-sample fit of the predictive regressions.

Following J. Z. Huang and Shi (2023), I also extract predictor factors from the large data set that contains the macroeconomic variables and their six lagged values. Table J.2 in the Appendix reports the coefficient estimates and statistics for these single predictor factors. The table also includes the results for the factor extraction methods other than PCA, sPCA, FLasso, GCALasso1 and GCALasso2. The table shows that, except for the FGCALasso1 and FGCALasso2 factors, the single predictor factors are significant at 5% according to both the conventional test and the bootstrap test in all cases.

Table J.3 in the Appendix reports the results regarding the adjusted R^2 for the five relevant methods in case the six lagged values of each macroeconomic variable are added to the data set. The table shows that adding lagged values only consistently improves the results in both samples for the GCALasso2 method. For this method, the adjusted R^2 increases from 25% to 72% in the earlier sample and from 16% to 74% in the later sample. The tremendous improvement in



(a) Restricted model with only the three yield PCs

(b) Model with the three PCs and GCALasso2 factor

Figure 3
In-sample fitted values for $\overline{rx}^{(5)}_{t+12}$ resulting from the restricted model with only the three yield curve PCs as predictors (on the left) and the unrestricted model with the three yield curve PCs and the GCALasso2 factor as predictors (on the right) along with the actual excess bond returns over the 1985-2022 sample. The GCALasso2 factor is extracted from the large macroeconomic data set that also includes the six lagged values of the macroeconomic variables. The adjusted R^2 for the restricted regression model is equal to 0.16 and the adjusted R^2 for the unrestricted regression model is equal to 0.74.

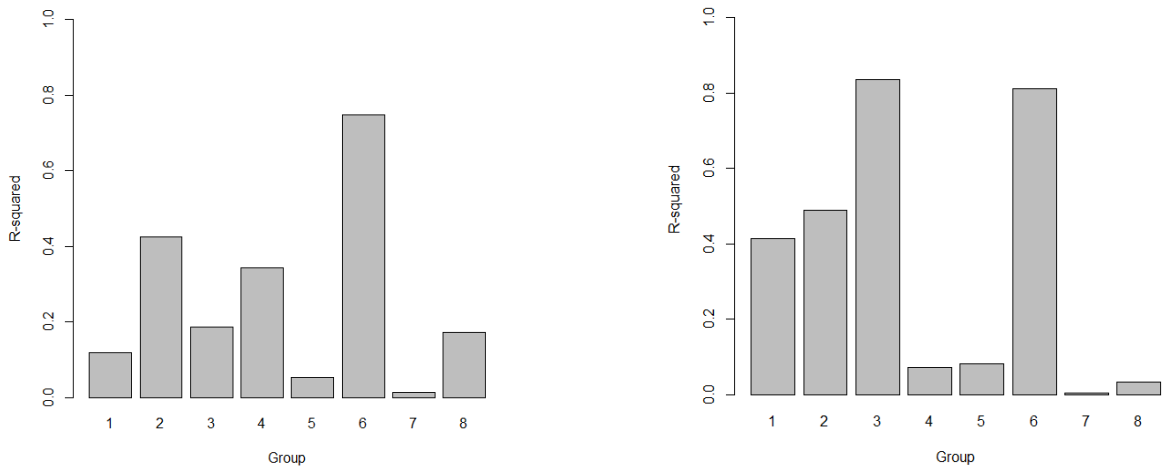
the in-sample fit in the later sample is also illustrated by Figure 3. This figure plots the fitted values for the excess bond returns resulting from the restricted and unrestricted model along with the actual values over the 1985-2022 sample. It can be seen that fitted values corresponding to the model including the GCALasso2 factor match the actual returns much more closely than the fitted values from the restricted model. A similar conclusion can be drawn from Figure J.1 which includes the plots over the 1964-2007 sample.

5.1.3 Economic Interpretation

Similar to Ludvigson and Ng (2009), I establish the economic interpretation of the predictor factors by looking at the R^2 of a regression of the factor on each macroeconomic variable contained in the large data set. In addition, I also regress the factor on all macroeconomic variables belonging to each of the eight groups specified in Table B.1 and deduce economic interpretations from the resulting R^2. I focus on the interpretation of the GCALasso2 factor extracted from the large macroeconomic data set including lagged values, as it is demonstrated that this factor has the largest in-sample predictive power for bond returns beyond the three yield PCs in both the earlier and later samples.

Figure 4 contains two bar plots, each displaying the R^2 statistics of the regressions of the GCALasso2 factor on the macroeconomic variables in each group. The first bar plot shows the results for the 1964-2007 sample, while the second shows the results for the 1985-2022 sample.

The figure shows that the explanatory power for the GCALasso2 factor of groups 2, 5, 6 and 7 is quite stable across the two samples. In both samples, the R^2 corresponding to group 6 is large and is equal to about 80%. Since group 6 consists of bond and FX variables, one may be concerned that the GCALasso2 factor is spanned by the yield curve. However, in Section 5.1.2 it is already shown that the GCALasso2 factor has significant in-sample predictive power beyond the three yield PCs. Apart from that, I also examine the validity of this concern using the informal test proposed by Duffee (2013). In particular, Duffee (2013) argues that if a variable is spanned by the yield curve a regression of this variable on the yield curve factors should produce serially uncorrelated fitted residuals. However, regressing the GCALasso2 factor on the three yield PCs leads to residuals with first-order autocorrelations equal to 0.66 in the earlier sample and 0.73 in the later sample. The autocorrelations are highly significant in both samples according to the Breusch-Godfrey test. This also suggests that the GCALasso2 factor is not spanned by the yield curve.



(a) 1964-2007 **(b)** 1985-2022

Figure 4
R^2 of the regressions of the in-sample GCALasso2 predictor factor on the macroeconomic variables in each of the eight groups estimated over the 1964-2007 sample (on the left) and the 1985-2022 sample (on the right). The eight groups are (1) output; (2) employment; (3) housing; (4) orders and inventories; (5) money market; (6) bond and foreign exchange (FX) market; (7) price indices; and (8) stock market. Table B.1 specifies which macroeconomic variables are included in each group. The GCALasso2 factor is constructed from the large data set containing the macroeconomic variables and their six lagged values.

The explanatory power of groups 1, 3, 4 and 8 is less stable across the two samples. Strikingly, the R^2 corresponding to the third group, which represents the housing factors, increases from about 19% in the 1964-2007 sample to 84% in the 1985-2022 sample. Together with the sixth group, the R^2 in the later sample is even 97%. This does not necessarily imply that the GCALasso2 factor in this sample mainly consists of variables from groups 3 and 6, as it could also contain variables of other groups (e.g., groups 1 and 2) that are correlated with variables

in groups 3 and 6. Nevertheless, the variables from these groups are able to explain almost all variation in the GCALasso2 factor and this means that those variables jointly contain the predictive power that is present in the GCALasso2 factor.

The explanatory dominance of the housing factors in the later sample is also clearly visible in Figure J.3 in the Appendix, which plots the R^2 statistics for the regression of the GCALasso2 factor on each macroeconomic variable as bar charts. The figure shows that the R^2 corresponding to 7 out of 10 housing factors is above 40%, whereas in the other groups none of the variables is able to individually explain such a large part of the variation in the GCALasso2 factor. Moreover, the explanatory power of the large majority of variables is even lower than 5%.

This is an interesting finding as the housing factors have been largely overlooked in the term structure literature (J. Z. Huang & Shi, 2023). Other groups have received much more attention. For example, employment and inflation variables, which are contained in groups 2 and 7, are commonly incorporated in macro-finance term structure models and are well motivated by for instance the equilibrium term structure model of Wachter (2006). A possible explanation for how the housing factors can be linked to bond risk premia is given by Piazzesi, Schneider and Tuzel (2007). They consider a consumption-based asset pricing model and show that the equity risk premium is driven by the expenditure share on housing.

5.2 Out-of-Sample Analysis

In the out-of-sample analysis, the forecasts resulting from the restricted model with only the three yield PCs are compared with the forecasts generated by the unrestricted model that also contains the additional predictors x_{2t}. As mentioned before, both the individual factors and the single predictor factor resulting from the factor extraction methods are considered as additional predictors x_{2t}. In the LN application, the single predictor factor is defined as the fitted value of the regression of $\overline{rx}^{(5)}_{t+12}$ on the eight macro PCs. The out-of-sample results corresponding to LN and the other relevant factor extraction methods for the different winsorization regimes are given in Table 5.

Firstly, I focus on the case in which no winsorization is applied. These results are reported in the first panel of the table. The results for LN are ambiguous. When adding either the eight macro PCs or the single predictor factor to the restricted model, the out-of-sample MSE improves by about 20% in the 2008-2016 sample, but it deteriorates in the 2008-2022 sample. In no case the change is statistically significant. Plot (d) in Figure K.1 in the Appendix demonstrates that the out-of-sample results for the 2008-2022 sample period are heavily impacted by outliers. In particular, extreme values for the macro PCs at the start of the COVID-19 pandemic in 2020

Table 5
Out-of-sample predictive power for $\overline{rx}_{t+12}^{(5)}$ of a restricted model with the three yield PCs and an unrestricted model with additional predictors x_{2t} corresponding to the LN application and the other relevant factor extraction methods in different scenarios. The results are shown for the scenarios in which no winsorization, input winsorization or output winsorization is applied. The additional predictors x_{2t} for the LN application are either the eight macro PCs or a single predictor factor which is the fitted value of the regression of $\overline{rx}_{t+12}^{(5)}$ on the eight macro PCs. The additional predictors x_{2t} for the PCA, sPCA and FLasso methods are either the six factors or the single predictor factor resulting from these methods. Since the GCALasso1 and GCALasso2 methods only produce a single predictor factor, the additional predictor x_{2t} for these methods is the factor resulting from each of these two methods. The in-sample period starts in January 1964 and ends in December 2007. The out-of-sample period starts one month later than the end of the in-sample period and ends in either December 2016 or December 2022. To generate the out-of-sample forecasts, expanding window estimation is used. Under *MSE ratio* and *p-value* the table reports respectively the mean-squared errors for the unrestricted model relative to the mean-squared errors for the restricted model and the p-values of the Diebold-Mariano test for equal prediction accuracy of the two models. The p-values that are lower than 5% are highlighted in bold.

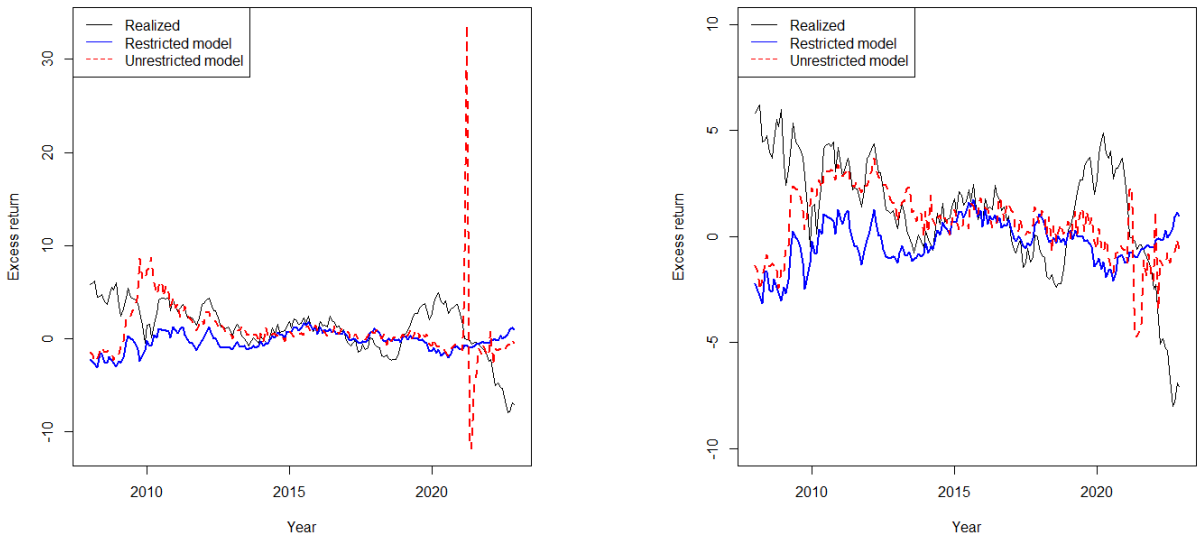
	Individual Factors				Single Joint Factor			
	End: 2016		End: 2022		End: 2016		End: 2022	
	MSE ratio	p-value	MSE ratio	p-value	MSE ratio	p-value	MSE ratio	p-value
No winsorization								
LN	0.778	0.279	1.844	0.437	0.814	0.416	1.259	0.602
PCA	0.753	0.171	1.098	0.786	0.776	0.231	1.431	0.526
sPCA	0.821	0.389	1.444	0.486	0.809	0.664	1.047	0.891
FLasso	0.861	0.620	1.250	0.562	0.894	0.742	1.329	0.489
GCALasso1					1.333	0.167	1.909	0.166
GCALasso2					1.442	0.225	2.021	0.194
Input winsorization								
LN	0.612	0.106	0.675	0.052	0.640	0.152	0.661	0.050
PCA	0.584	0.054	0.656	**0.024**	0.617	0.062	0.647	**0.018**
sPCA	0.701	0.083	0.733	0.058	0.523	0.113	0.618	0.061
FLasso	0.466	0.109	0.600	0.061	0.579	0.124	0.690	0.073
GCALasso1					1.256	0.188	1.108	0.521
GCALasso2					1.364	0.169	1.168	0.386
Output winsorization								
LN	0.612	0.106	0.675	0.052	0.640	0.152	0.661	0.050
PCA	0.560	0.073	0.643	**0.030**	0.572	0.051	0.673	**0.037**
sPCA	0.577	0.069	0.671	**0.041**	0.438	0.130	0.609	0.086
FLasso	0.527	0.059	0.652	**0.037**	0.527	**0.039**	0.702	0.058
GCALasso1					1.152	0.521	1.173	0.303
GCALasso2					1.145	0.581	1.103	0.546

cause the unrestricted model to generate some extreme 12-months ahead forecasts for 2021.

The out-of-sample results for PCA, sPCA and FLasso are similar to those for LN. Adding either the six individual factors or the single predictor factor resulting from these methods to the restricted forecasting model improves the out-of-sample MSE in the 2008-2016 sample, but leads to a deterioration in the 2008-2022 sample. Including the GCALasso1 or GCALasso2 factor in the restricted model results in higher prediction errors in both samples.

Like in the LN application, results are severely affected by outliers. Therefore, it is sensible to use the winsorization methods in order to reduce the impact of these extreme values. The out-of-sample results for the scenarios in which either input or output winsorization is applied are given in the second and third panels respectively. It can be seen that the out-of-sample performance of all methods improves substantially, especially in the 2008-2022 sample. This is also illustrated in Figure 5 which plots the forecasts resulting from the restricted and unrestricted model with the single predictor factor extracted using the PCA method both in case no winsorization and in case output winsorization is applied.

For both winsorization methods, the reduction in the MSE as a result of adding either the six



(a) No winsorization **(b)** Output winsorization

Figure 5
Out-of-sample forecasts for $\overline{rx}_{t+12}^{(5)}$ resulting from the restricted forecast model with only the three yield curve PCs as predictors and the unrestricted model with the three yield curve PCs and the single predictor factor extracted using the PCA method in case no winsorization is applied (on the left) and in case output winsorization is applied (on the right). Expanding window estimation is used to construct the forecasts. The training windows start in January 1964. The out-of-sample period starts in January 2008 and ends in December 2022.

individual factors or the single predicor factor resulting from the PCA method to the restricted model is significant at 10% level in the 2008-2016 sample and even at 5% in the 2008-2022 sample. The improvements in MSE are quite similar for the sPCA and FLasso methods but are significant at 10% in the 2008-2022 sample only. Winsorization also helps to improve the out-of-sample performance of the forecasting model including the GCALasso1 or GCALasso2 factor in addition to the three yield PCs. Nevertheless, the performance of this model remains inferior to the restricted forecasting model.

Table K.1 in the Appendix also documents the out-of-sample results corresponding to all the other methods and for the scenarios in which the factors are extracted from the large macroeconomic data that includes the lagged values. Overall, results are similar regardless of whether input winsorization or output winsorization is applied and whether the individual factors or the single predictor factors are used as additional predictors x_{2t}. Furthermore, most methods produce predictor factors that are able to considerably improve the restricted forecasting model. However, the improvements are often not significant at 10%. The PCA method leads to the most robust and consistent results. In case winsorization is applied, the factors resulting from this method improve the MSE by at least 38% in the 1964-2007 sample and 32% in the 1985-2022 sample. Moreover, all these MSE improvements are significant at 10% level. A number of methods lead to better results in certain scenarios and sample periods, but none of the methods consistently outperforms the PCA method. Adding lagged values is beneficial for some methods,

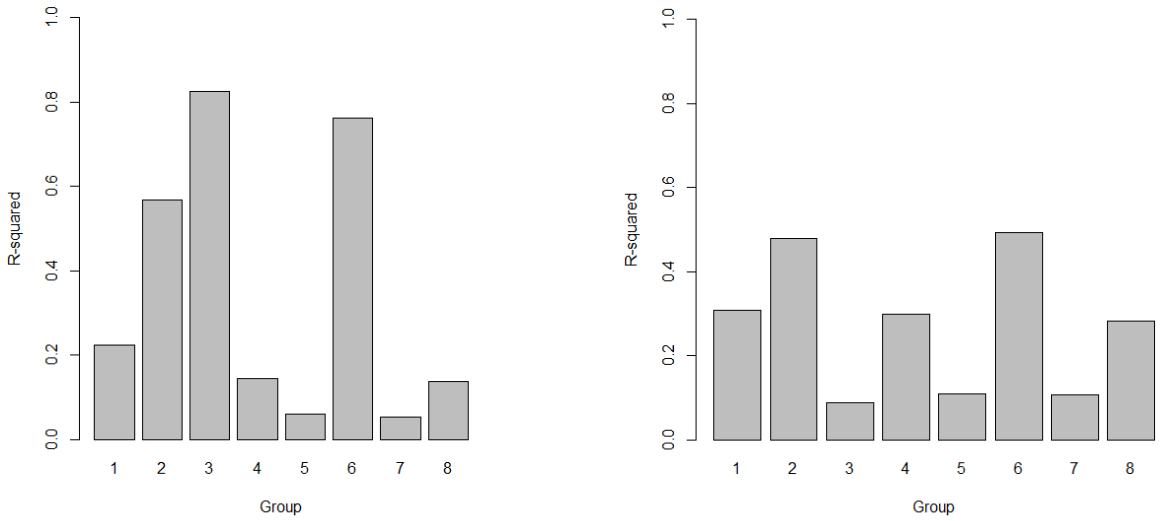
including the GCALasso2 method, but it does not change conclusions. It is interesting to note that despite achieving the best in-sample results, the GCALasso2 method leads to the worst out-of-sample results, even in scenarios with winsorization and lagged macroeconomic variables. A potential explanation for this is given in the next section.

5.2.1 Economic Interpretation

Similar to the in-sample analysis, I also interpret the most relevant out-of-sample single predictor factors economically. I focus on the interpretation of the factor resulting from the PCA method, as the out-of-sample predictive power of this factor is consistently significant. Furthermore, I interpret the GCALasso2 factor that is extracted from the large macroeconomic data set with lagged values. This factor is interesting as it has the largest in-sample but also the smallest out-of-sample predictive power for excess bond returns beyond the three yield PCs compared to the factors of the other methods. Output winsorization is applied to both factors. Similar results are obtained in the case of input winsorization.

The R^2 statistics for the regression of these factors on the macroeconomic variables in each of the eight groups estimated over the 1985-2022 sample are depicted in Figure 6. Interestingly, the pattern of the bar plot for the out-of-sample PCA factor is comparable to the bar plot for the in-sample GCALasso2 factor shown in Figure 4. The housing group has the largest explanatory power for the PCA factor, followed by the bond and FX group, and then the employment group. The groups explain respectively 83%, 76% and 57% of the variation in the PCA factor. Jointly, these groups are able to explain almost all variation in the PCA factor. From Figure K.2, which displays the R^2 statistics corresponding to the individual macroeconomic variables, it can also be seen that the housing factors are the most prominent factors in terms of explanatory power.

The out-of-sample bar plot for the GCALasso2 factor is quite similar to its in-sample plot for the 1964-2007 sample. Since in Section 5.1.3 it has been shown that the composition of the in-sample GCALasso2 factor is quite different in the 1964-2007 sample compared to the 1985-2022 sample, this is an indication that the out-of-sample factor is heavily impacted by the inclusion of the data from 1964 through 1984 in the training windows. However, the change in the composition of the in-sample GCALasso2 factor also implies that the predictive ability of certain macroeconomic variables possibly changed over time. The information in the 1964-1984 sample may not be relevant anymore for forecasting bond returns for the 2008-2022 sample. Including the potentially outdated data in the training windows may influence the out-of-sample GCALasso2 factor such that it is less responsive to changes in the data and less useful for the prediction of excess bond returns.



(a) PCA factor

(b) GCALasso2 factor

Figure 6
R^2 of the regressions of the out-of-sample PCA single predictor factor (on the left) and the GCALasso2 predictor factor (on the right) on the macroeconomic variables in each of the eight groups. The out-of-sample period is The eight groups are (1) output; (2) employment; (3) housing; (4) orders and inventories; (5) money market; (6) bond and foreign exchange (FX) market; (7) price indices; and (8) stock market. Table B.1 specifies which macroeconomic variables are included in each group. The predictor factors are obtained using expanding window estimation with training windows starting in January 1964 and output winsorization is recursively applied to the resulting factors. The regressions are estimated using data from January 2007 to December 2021. The factors constructed for this sample are namely used to produce out-of-sample forecasts for $\overline{rx}^{(5)}_{t+12}$ for the 2008-2022 period. The PCA factor is extracted from the large macroeconomic data set without lagged values, whereas the GCALasso2 factor is obtained from the large data set with lagged values.

5.2.2 Robustness Checks

I also investigate the robustness of the out-of-sample results for the most relevant methods with respect to the number of predictor factors and the winsorization thresholds. The out-of-sample results for different numbers of factors and winsorization thresholds are given in Table K.2 and K.3 respectively. These results are analysed in detail in Appendix K.4. Overall, it can be concluded that the results are robust to alternative numbers of predictor factors and winsorization thresholds.

6 Conclusion

Based on the simulation study and empirical analysis, I confirm the finding of Bauer and Hamilton (2018) that conventional tests are unreliable for inference about the spanning hypothesis as they suffer from serious small-sample econometric problems. The bootstrap test proposed by Bauer and Hamilton (2018) is robust to these problems and using this test to revisit the evidence of four widely cited studies leads to weaker evidence against the spanning hypothesis than suggested in these studies. Moreover, adding new data and evaluating the out-of-sample predictive power for excess bond returns beyond the three yield principal components

of the proposed additional variables leads to evidence that is even far from convincing. These findings are in line with Bauer and Hamilton (2018) and reinforce the spanning controversy in macro-finance literature.

However, after extending the study of Ludvigson and Ng (2009) by using alternative methods to extract predictor factors from a large set of macroeconomic variables and using winsorization methods to reduce the impact of extreme values, I resolve this spanning controversy. The in-sample predictive power above the three yield PCs of the single predictor factor resulting from these alternative methods, including 12 machine-learning methods, is evaluated over an earlier and a later sample period. With a few exceptions, the results for all methods are highly significant according to the bootstrap test in both samples and even in scenarios without winsorization. I find that the GCALasso2 factor that is extracted from the macroeconomic data set with lagged values has the strongest in-sample additional predictive power for excess bond returns. Adding it to the restricted regression model with only the three yield PCs leads to a tremendous increase in the adjusted R^2 and this increase is highly significant. Due to extreme values in the large macroeconomic data set, winsorization is crucial for obtaining good out-of-sample forecasts when using factors extracted from this data set as predictors. In case winsorization is applied, I find that the performance of the PCA method is the most consistent and robust. It produces predictor factors with significant predictive power beyond the first three yield curve principal components both in-sample and out-of-sample and across different sample periods. Despite the good in-sample performance, the GCALasso1 and GCALasso2 methods display the poorest out-of-sample performance among all methods.

Another interesting finding is that in the recent samples the housing variables play a prominent role in both the best-performing in-sample and out-of-sample predictor factors. In fact, the explanatory power of the variables in the housing group is larger than 80% for both predictor factors and is the largest compared to the explanatory power of the other groups. It seems that at least part of the incremental predictive power for excess bond returns over the yield PCs in the post-1985 samples can be attributed to these variables. However, this implication is not formally tested and the precise amount of incremental predictive power in these variables is not quantified. This would be an interesting avenue for future research. Furthermore, it would be interesting to construct real-time implementable trading strategies using the best-performing factor extraction methods and investigate whether these can generate significant economic gains for investors.

References

Ang, A. & Bekaert, G. (2006). Stock return predictability: Is it there? *Review of Financial Studies*, *20*(3), 651-707. doi: 10.1093/rfs/hhl021

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

Bottmer, L., Croux, C. & Wilms, I. (2022). Sparse regression for large data sets with outliers. *European Journal of Operational Research*, *297*(2), 782–794.

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

Cavanagh, C. L., Elliott, G. & Stock, J. H. (1995). Inference in models with nearly integrated regressors. *Econometric Theory*, *11*(5), 1131-1147. doi: 10.1017/s0266466600009981

Chan, N. H. (1988). The parameter inference for nearly nonstationary time series. *Journal of the American Statistical Association*, *83*(403), 857-862. doi: 10.1080/01621459.1988.10478674

Cieslak, A. & Povala, P. (2015). Expected returns in treasury bonds. *Review of Financial Studies*, *28*(10), 2859-2901. doi: 10.1093/rfs/hhv032

Cochrane, J. H. & Piazzesi, M. (2005). Bond risk premia. *American Economic Review*, *95*(1), 138-160. doi: 10.1257/0002828053828581

Cooper, I. & Priestley, R. (2008). Time-varying risk premiums and the output gap. *Review of Financial Studies*, *22*(7), 2801-2833. doi: 10.1093/rfs/hhn087

Diebold, F. X. & Mariano, R. S. (2002). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, *20*(1), 134-144. doi: 10.1198/073500102753410444

Duffee, G. R. (2013). Chapter 13 bond pricing and the macroeconomy. In (p. 907-967). doi: 10.1016/b978-0-44-459406-8.00013-5

Greenwood, R. & Vayanos, D. (2014). Bond supply and excess bond returns. *Review of Financial Studies*, *27*(3), 663-713. doi: 10.1093/rfs/hht133

Gürkaynak, R. S., Sack, B. & Wright, J. H. (2007). The u.s. treasury yield curve: 1961 to the present. *Journal of Monetary Economics*, *54*(8), 2291-2304. doi: 10.1016/j.jmoneco.2007.06.029

Gürkaynak, R. S. & Wright, J. H. (2012). Macroeconomics and the term structure. *Journal of Economic Literature*, *50*(2), 331-367. doi: 10.1257/jel.50.2.331

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

Hodrick, R. J. (1992). Dividend yields and expected stock returns: Alternative procedures for

inference and measurement. *Review of Financial Studies*, *5*(3), 357-386. doi: 10.1093/rfs/5.3.351

Hoerl, A. E. & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67.

Huang, D., Jiang, F., Li, K., Tong, G. & Zhou, G. (2022). Scaled pca: A new approach to dimension reduction. *Management Science*, *68*(3), 1678–1695.

Huang, D., Jiang, F., Li, K., Tong, G. & Zhou, G. (2023). Are bond returns predictable with real-time macro data? *Journal of Econometrics*. doi: 10.1016/j.jeconom.2022.09.008

Huang, J. Z. & Shi, Z. (2023). Machine-learning-based return predictors and the spanning controversy in macro-finance. *Management Science*, *69*(3), 1780-1804. doi: 10.1287/mnsc.2022.4386

Joslin, S., Priebsch, M. & Singleton, K. J. (2014). Risk premiums in dynamic term structure models with unspanned macro risks. *The Journal of Finance*, *69*(3), 1197-1233. doi: 10.1111/jofi.12131

Kilian, L. (1998). Small-sample confidence intervals for impulse response functions. *The Review of Economics and Statistics*, *80*(2), 218-230. doi: 10.1162/003465398557465

Le, A. & Singleton, K. J. (2013). The structure of risks in equilibrium affine models of bond yields. *Unpublished working paper, University of North Carolina at Chapel Hill*.

Litterman, R. B. & Scheinkman, J. (1991). Common factors affecting bond returns. *The Journal of Fixed Income*, *1*(1), 54-61. doi: 10.3905/jfi.1991.692347

Ludvigson, S. C. & Ng, S. (2009). Macro factors in bond risk premia. *Review of Financial Studies*, *22*(12), 5027-5067. doi: 10.1093/rfs/hhp081

Ludvigson, S. C. & Ng, S. (2016). A factor analysis of bond risk premia. In *Handbook of empirical economics and finance* (p. 313-371). doi: 10.1201/b10440-13

Newey, W. K. & West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, *55*(3), 703. doi: 10.2307/1913610

Phillips, P. C. B. (1988). Regression theory for near-integrated time series. *Econometrica*, *56*(5), 1021. doi: 10.2307/1911357

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

Pope, A. L. (1990). Biases of estimators in multivariate non-gaussian autoregressions. *Journal of Time Series Analysis*, *11*(3), 249-258. doi: 10.1111/j.1467-9892.1990.tb00056.x

Stambaugh, R. F. (1999). Predictive regressions. *Journal of Financial Economics*, *54*(3),

375-421. doi: 10.1016/s0304-405x(99)00041-0

Stock, J. H. (1991). Confidence intervals for the largest autoregressive root in u.s. macroeconomic time series. *Journal of Monetary Economics*, *28*(3), 435-459. doi: 10.1016/0304-3932(91)90034-l

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

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

Wei, M. & Wright, J. H. (2011). Reverse regressions and long-horizon forecasting. *Journal of Applied Econometrics*, *28*(3), 353-371. doi: 10.1002/jae.1274

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

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

Appendix A Variables used in JPS, CPO and CP

Table 1 shows that the variables used in the research of Joslin et al. (2014) are economic growth, measured by the 3-month moving average Chicago Fed National Activity Index, inflation, defined as the 1-year inflation predicted by the Blue Chip Financial Forecasts, and bond yields, constructed by Anh Le (Le & Singleton, 2013) using the Fama-Bliss selection criteria. Similar to Ludvigson and Ng (2009), Cochrane and Piazzesi (2005) use data on the bond prices of pure discount U.S Treasury bonds with maturities ranging from one to five years. These data are obtained from the Fama-Bliss dataset from the Center for Research in Securities Prices (CRSP). To replicate and extend the paper of Cieslak and Povala (2015), Bauer and Hamilton (2018) use the Consumer Price Index, retrieved from the FRED database, the 1-month T-bill rate, obtained from the CRSP, and the zero-coupon yields with maturities ranging from one to fifteen years, constructed by Gürkaynak, Sack and Wright (2007). For more details on the construction of the variables, I refer to Bauer and Hamilton (2018) and the corresponding revisited paper.

Appendix B Description of Large Macroeconomic Data Set

Table B.1
Data description of the extended large macroeconomic data set. The series number, the short name, and the full name of each variable are specified in the first, second and third columns respectively. The fourth column indicates to which of the eight groups each variable belongs: (1) output; (2) employment; (3) housing; (4) orders and inventories; (5) money market; (6) bond and foreign exchange (FX) market; (7) price indices; and (8) stock market. The last column indicates which transformation is applied to each variable: (1) no transformation; (2) first difference; (3) second difference; (4) logarithm; (5) first difference of logarithm; and (6) second difference of logarithm.

ID	Short name	Description	Group	Transformation
1	PI	Real Personal Income	1	5
2	PI less transfers	Real personal income ex transfer receipts	1	5
3	Real Consumption	Real personal consumption expenditures	4	5
4	M&T sales	Real Manu. and Trade Industries Sales	4	5
5	Retail sales	Retail and Food Services Sales	4	5
6	IP: total	IP Index	1	5
7	IP: products	IP: Final Products and Nonindustrial Supplies	1	5
8	IP: final prod	IP: Final Products (Market Group)	1	5
9	IP: cons gds	IP: Consumer Goods	1	5
10	IP: cons dble	IP: Durable Consumer Goods	1	5
11	IP: cons nondble	IP: Nondurable Consumer Goods	1	5
12	IP: bus eqpt	IP: Business Equipment	1	5
13	IP: matls	IP: Materials	1	5
14	IP: dble matls	IP: Durable Materials	1	5
15	IP: nondble matls	IP: Nondurable Materials	1	5
16	IP: mfg	IP: Manufacturing (SIC)	1	5
17	IP: res util	IP: Residential Utilities	1	5
18	IP: fuels	IP: Fuels	1	5
20	Cap util	Capacity Utilization: Manufacturing	1	2
21	Help wanted indx	Help-Wanted Index for United States	2	2
22	Help wanted/une	Ratio of Help Wanted/No. Unemployed	2	2
23	Emp CPS total	Civilian Labor Force	2	5
24	Emp CPS nonag	Civilian Employment	2	5
25	U: all	Civilian Unemployment Rate	2	2
26	U: mean duration	Average Duration of Unemployment (Weeks)	2	2
27	U $<$ 5 wks	Civilians Unemployed - Less Than 5 Weeks	2	5
28	U 5-14 wks	Civilians Unemployed for 5-14 Weeks	2	5
29	U 15+ wks	Civilians Unemployed - 15 Weeks & Over	2	5
30	U 15-26 wks	Civilians Unemployed for 15-26 Weeks	2	5
31	U 27+ wks	Civilians Unemployed for 27 Weeks and Over	2	5
32	UI claims	Initial Claims	2	5
33	Emp: total	All Employees: Total nonfarm	2	5
34	Emp: gds prod	All Employees: Goods-Producing Industries	2	5
35	Emp: mining	All Employees: Mining and Logging: Mining	2	5
36	Emp: const	All Employees: Construction	2	5
37	Emp: mfg	All Employees: Manufacturing	2	5
38	Emp: dble gds	All Employees: Durable goods	2	5
39	Emp: nondbles	All Employees: Nondurable goods	2	5
40	Emp: services	All Employees: Service-Providing Industries	2	5
41	Emp: TTU	All Employees: Trade, Transportation & Utilities	2	5
42	Emp: wholesale	All Employees: Wholesale Trade	2	5
43	Emp: retail	All Employees: Retail Trade	2	5
44	Emp: FIRE	All Employees: Financial Activities	2	5
45	Emp: Govt	All Employees: Government	2	5
46	Avg hrs	Avg Weekly Hours : Goods-Producing	2	1
47	Overtime: mfg	Avg Weekly Overtime Hours : Manufacturing	2	2
48	Avg hrs: mfg	Avg Weekly Hours : Manufacturing	2	1
50	Starts: nonfarm	Housing Starts: Total New Privately Owned	3	4
51	Starts: NE	Housing Starts, Northeast	3	4
52	Starts: MW	Housing Starts, Midwest	3	4
53	Starts: South	Housing Starts, South	3	4
54	Starts: West	Housing Starts, West	3	4
55	BP: total	New Private Housing Permits (SAAR)	3	4
56	BP: NE	New Private Housing Permits, Northeast (SAAR)	3	4
57	BP: MW	New Private Housing Permits, Midwest (SAAR)	3	4
58	BP: South	New Private Housing Permits, South (SAAR)	3	4
59	BP: West	New Private Housing Permits, West (SAAR)	3	4
64	Orders: cons gds	New Orders for Consumer Goods	4	5
65	Orders: dble gds	New Orders for Durable Goods	4	5

Continued on next page

Continued from previous page

ID	Short name	Description	Group	Transformation
66	Orders: cap gds	New Orders for Nondefense Capital Goods	4	5
67	Unf orders: dble	Unfilled Orders for Durable Goods	4	5
68	M&T invent	Total Business Inventories	4	5
69	M&T invent/sales	Total Business: Inventories to Sales Ratio	4	2
70	M1	M1 Money Stock	5	6
71	M2	M2 Money Stock	5	6
72	M2 (real)	Real M2 Money Stock	5	5
73	MB	Monetary Base	5	6
74	Reserves tot	Total Reserves of Depository Institutions	5	6
75	Reserves nonbor	Reserves Of Depository Institutions	5	7
76	C&I loan plus	Commercial and Industrial Loans	5	6
77	DC&I loans	Real Estate Loans at All Commercial Banks	5	6
78	Cons credit	Total Nonrevolving Credit	5	6
79	Inst cred/PI	Nonrevolving consumer credit to Personal Income	5	2
80	S&P 500	S&P's Common Stock Price Index: Composite	8	5
81	S&P: indust	S&P's Common Stock Price Index: Industrials	8	5
82	S&P div yield	S&P's Composite Common Stock: Dividend Yield	8	2
83	S&P PE ratio	S&P's Composite Common Stock: Price-Earnings Ratio	8	5
84	Fed Funds	Effective Federal Funds Rate	6	2
85	Comm paper	3-Month AA Financial Commercial Paper Rate	6	2
86	3 mo T-bill	3-Month Treasury Bill:	6	2
87	6 mo T-bill	6-Month Treasury Bill:	6	2
88	1 yr T-bond	1-Year Treasury Rate	6	2
89	5 yr T-bond	5-Year Treasury Rate	6	2
90	10 yr T-bond	10-Year Treasury Rate	6	2
91	Aaa bond	Moody's Seasoned Aaa Corporate Bond Yield	6	2
92	Baa bond	Moody's Seasoned Baa Corporate Bond Yield	6	2
93	CP-FF spread	3-Month Commercial Paper Minus FEDFUNDS	6	1
94	3 mo-FF spread	3-Month Treasury C Minus FEDFUNDS	6	1
95	6 mo-FF spread	6-Month Treasury C Minus FEDFUNDS	6	1
96	1 yr-FF spread	1-Year Treasury C Minus FEDFUNDS	6	1
97	5 yr-FF spread	5-Year Treasury C Minus FEDFUNDS	6	1
98	10 yr-FF spread	10-Year Treasury C Minus FEDFUNDS	6	1
99	Aaa-FF spread	Moody's Aaa Corporate Bond Minus FEDFUNDS	6	1
100	Baa-FF spread	Moody's Baa Corporate Bond Minus FEDFUNDS	6	1
101	Ex rate: avg	Trade Weighted U.S. Dollar Index	6	5
102	Ex rate: Switz	Switzerland / U.S. Foreign Exchange Rate	6	5
103	Ex rate: Japan	Japan / U.S. Foreign Exchange Rate	6	5
104	Ex rate: UK	U.S. / U.K. Foreign Exchange Rate	6	5
105	EX rate: Canada	Canada / U.S. Foreign Exchange Rate	6	5
106	PPI: fin gds	PPI: Finished Goods	7	6
107	PPI: cons gds	PPI: Finished Consumer Goods	7	6
108	PPI: int matls	PPI: Intermediate Materials	7	6
109	PPI: crude matls	PPI: Crude Materials	7	6
110	Spot market price	Crude Oil, spliced WTI and Cushing	7	6
111	PPI: nonferrous	PPI: Metals and metal products:	7	6
113	CPI-U: all	CPI : All Items	7	6
114	CPI-U: apparel	CPI : Apparel	7	6
115	CPI-U: transp	CPI : Transportation	7	6
116	CPI-U: medical	CPI : Medical Care	7	6
117	CPI-U: comm.	CPI : Commodities	7	6
118	CPI-U: dbles	CPI : Durables	7	6
119	CPI-U: services	CPI : Services	7	6
120	CPI-U: ex food	CPI : All Items Less Food	7	6
121	CPI-U: ex shelter	CPI : All items less shelter	7	6
122	CPI-U: ex med	CPI : All items less medical care	7	6
123	PCE defl	Personal Cons. Expend.: Chain Index	7	6
124	PCE defl: dlbes	Personal Cons. Exp: Durable goods	7	6
125	PCE defl: nondble	Personal Cons. Exp: Nondurable goods	7	6
126	PCE defl: service	Personal Cons. Exp: Services	7	6
127	AHE: goods	Avg Hourly Earnings : Goods-Producing	2	6
128	AHE: const	Avg Hourly Earnings : Construction	2	6
129	AHE: mfg	Avg Hourly Earnings : Manufacturing	2	6
130	Consumer expect	Consumer Sentiment Index	4	2
132	N.A.	Consumer Motor Vehicle Loans Outstanding	5	6
133	N.A.	Total Consumer Loans and Leases Outstanding	5	6
134	N.A.	Securities in Bank Credit at All Commercial Banks	5	6
135		VIX	8	1

Appendix C Variables in Predictive Regressions for JPS, CPO and CP

The variables used in the predictive regressions to revisit JPS, CPO, and CP are described in this section. For these applications, I use the same variables and data as Bauer and Hamilton (2018) employ to revisit these studies. Since the target variable y_{t+h}, the bond yields i_t used to extract the first three PCs for x_{1t}, and the additional predictors x_{2t} differ across these applications, an overview of the variables used in each application is provided in Table C.1.

Table C.1
The dependent variable (y_{t+h}), the bond yields (i_t) from which the first three PCs are extracted to construct x_{1t}, the additional predictors (x_{2t}), the autocorrelation correction (AC) method, and the bootstrap method that are used in this paper to revisit the evidence of JPS, CPO and CP. The dependent variable is specified according to Equation (6) or (12). The additional predictors are described in the text. The autocorrelation method is either the approach in which the Newey-West (NW) standard errors with 18 lags are used or the reverse-regression (RR) approach of Wei and Wright (2011). Both the simple and the bias-corrected bootstrap methods are described in Appendix F.

Application	y_{t+12}	i_t	x_{2t}	AC method	Bootstrap method
JPS	$\overline{rx}^{(10)}_{t+12}$	$(i_t^{(6)}, i_t^{(12)}, i_t^{(24)}, i_t^{(36)}, ..., i_t^{(120)})'$	(GRO_t, INF_t)	NW	Bias-corrected bootstrap
CPO	$\overline{wrx}^{(15)}_{t+12}$	$(i_t^{(1)}, i_t^{(12)}, i_t^{(24)}, i_t^{(36)}, ..., i_t^{(180)})'$	τ_t	RR	Bias-corrected bootstrap
CP	$\overline{rx}^{(5)}_{t+12}$	$(i_t^{(12)}, i_t^{(24)}, i_t^{(36)}, i_t^{(48)}, i_t^{(60)})'$	$(PC_t^{(4)}, PC_t^{(5)})$	NW	Simple bootstrap

As described in Section A, the zero-coupon yields used in the JPS application are constructed by Anh Le and the bond yields used in the CP application are obtained from Gürkaynak et al. (2007). Similar to the LN application, the bond yields for the CP application are computed from the data on the bond prices of the zero-coupon Treasury bonds with maturities from one to five years using Equation (3). For the CPO application, the weighted average of excess bond returns is used instead of the unweighted average. The weighted average of annual excess bond returns across bonds with maturities ranging from 2 to k years is defined as

$$\overline{wrx}^{(k)}_{t+12} = \frac{1}{k-1}\sum_{n=2}^{k}\frac{1}{n}rx^{(12n)}_{t+12} \tag{12}$$

where $k \geq 2$.

Table C.1 also indicates which additional predictors x_{2t} are used in each application. The additional predictors used in the predictive regression to revisit each of the four earlier mentioned studies correspond to those used by Bauer and Hamilton (2018). From the table, it can be seen that in the case of the JPS application x_{2t} consists of the measure of economic growth (GRO) and inflation (INF) as specified in Section A. For CPO x_{2t} is set equal to the exponentially weighted moving average (EWMA) of the year-over-year inflation in month t, which is a measure of trend

inflation and is defined as

$$\tau_t = (1-\lambda)\sum_{j=0}^{t}\lambda^j \pi_{t-j} \tag{13}$$

where π_t is the year-over-year inflation in the CPI in month t and $\lambda = 0.987$. This value for λ leads to the strongest results in Cieslak and Povala (2015). To revisit CP, x_{2t} is equalized to the vector containing the fourth and fifth PCs of yields on bonds with maturities of one through five years. The values of these PCs in month t are denoted by $PC_t^{(4)}$ and $PC_t^{(5)}$.

Appendix D GCALasso1 and GCALasso2 Method

In the first step of the GCALasso1 and GCALasso2 methods, Controlled Adaptive Lasso (**CALasso**) is applied to the variables in each of the eight groups specified in Section 2 separately. The corresponding minimization problem for group j is formulated as

$$\min_{\beta^{(j)}} \left(\sum_{t=1}^{T-12} \left(\overline{rx}_{t+12}^{(5)} - \sum_{k=1}^{3} PC_t^{(k)} \beta_{1k}^{(j)} - \sum_{i=1}^{N_j} z_{t,j}^{(i)} \beta_{2i}^{(j)} \right)^2 + \sum_{i=1}^{N_j} \frac{\lambda}{|\hat{\beta}_{i,j,1}^{ridge}|} |\beta_{2i}^{(j)}| \right) \tag{14}$$

where $z_{t,j}^{(i)}$ is the ith variable in group j, N_j is the number of variables in group j, $\hat{\beta}_{i,j,1}^{ridge}$ is the estimated coefficient for variable i in group j of the first-stage ridge regression of $\overline{rx}_{t+12}^{(5)}$ on the macroeconomic variables in group j, and λ is determined using five-fold cross-validation. The set of macroeconomic variables in group j that do not have a zero coefficient in the first step is denoted by $\hat{Z}_j$, $j = 1, \ldots, 8$.

In the second step, all groups are considered together and a variant of the group Lasso proposed by Yuan and Lin (2006) is applied to the macroeconomic variables selected in the first step. In this variant, the tuning parameter is the same within the groups but differs across groups. Specifically, the minimization problem that is solved in this method, termed the **GCALasso1** method, is given by

$$\min_{\beta} \left(\sum_{t=1}^{T-12} \left(\overline{rx}_{t+12}^{(5)} - \sum_{k=1}^{3} PC_t^{(k)} \beta_{1k}^{(j)} - \sum_{j=1}^{8} \sum_{i=1}^{\hat{N}_j} \hat{z}_{t,j}^{(i)} \beta_{2i}^{(j)} \right)^2 + \sum_{j=1}^{8} \sum_{i=1}^{\hat{N}_j} \lambda^{(j)} |\beta_{2i}^{(j)}| \right) \tag{15}$$

where $\hat{z}_{t,j}^{(i)}$ denotes the ith macroeconomic variable in group j that survives the first stage, $\hat{N}_j$ denote the number of selected variables in group j, and $\lambda^{(j)}$ is defined as

$$\lambda^{(j)} = \frac{\lambda}{\sqrt{\sum_{i=1}^{\hat{N}_j} \left(\hat{\beta}_{i,j,2}^{ridge} \right)^2}} \tag{16}$$

where $\hat{\beta}_{i,j,2}^{ridge}$ is the estimated coefficient for variable i in group j of the second-stage ridge regression of $\overline{rx}_{t+12}^{(5)}$ on the all macroeconomic variables that survive the first stage and λ is obtained from five-fold cross-validation. Additionally, I also consider the variant, termed **GCALasso2**, in which the tuning parameter also varies within the groups. To this end, $\lambda^{(j)}$ in minimization problem (15) is replaced by $\lambda_i^{(j)}$ defined as

$$\lambda_i^{(j)} = \frac{\lambda}{\hat{\beta}_{i,j,2}^{ridge}}. \tag{17}$$

Appendix E Pseudocode for Machine-Learning Methods

Algorithm 1 Pseudocode to construct predictors using the machine-learning methods

Input Data: $\overline{rx}^{(5)}$, Z
Input Binary State Variables: $Control, Adaptive, Group, Factor \in \{true, false\}$, $Option \in \{1, 2\}$

if $Group = false$ **then**
  if $Adaptive = false$ **then**
    if $Control = false$ **then**
      Solve minimization problem (7) yielding $\hat{\beta}^{Lasso}$
      if $Factor = false$ **then**
        Lasso: Use $\sum_{i=1}^{N} \hat{\beta}_i^{Lasso} z^{(i)}$ as a single additional predictor
      else
        FLasso: Apply PCA as described in Section 3.3.1 to the scaled variables $(\hat{\beta}_1^{Lasso} z^{(1)}, \ldots, \hat{\beta}_N^{Lasso} z^{(N)})$ and use either the six resulting principal components or the single predictor factor as additional predictor(s)
      end if
    else
      Solve minimization problem (8) yielding $\hat{\beta}^{CLasso}$
      if $Factor = false$ **then**
        CLasso: Use $\sum_{i=1}^{N} \hat{\beta}_{2i}^{CLasso} z^{(i)}$ as a single additional predictor
      else
        FCLasso: Apply PCA as described in Section 3.3.1 to the scaled variables $(\hat{\beta}_{21}^{CLasso} z^{(1)}, \ldots, \hat{\beta}_{2N}^{CLasso} z^{(N)})$ and use either the six resulting principal components or the components or the single predictor factor as additional predictor(s)
      end if
    end if
  else
    if $Control = false$ **then**
      Solve minimization problem (7) where λ is replaced by λ_i defined in Equation (9) with $\gamma = 1$ yielding $\hat{\beta}^{ALasso}$
      if $Factor = false$ **then**
        ALasso: Use $\sum_{i=1}^{N} \hat{\beta}_i^{ALasso} z^{(i)}$ as a single additional predictor
      else
        FALasso: Apply PCA as described in Section 3.3.1 to the scaled variables $(\hat{\beta}_1^{ALasso} z^{(1)}, \ldots, \hat{\beta}_N^{ALasso} z^{(N)})$ and use either the six resulting principal components or the single predictor factor as additional predictor(s)
      end if
    else
      Solve minimization problem (8) where λ is replaced by λ_i defined in Equation (9) with $\gamma = 1$ yielding $\hat{\beta}^{CALasso}$
      if $Factor = false$ **then**
        CALasso: Use $\sum_{i=1}^{N} \hat{\beta}_i^{CALasso} z^{(i)}$ as a single additional predictor
      else
        FCALasso: Apply PCA as described in Section 3.3.1 to the scaled variables $(\hat{\beta}_1^{CALasso} z^{(1)}, \ldots, \hat{\beta}_N^{CALasso} z^{(N)})$ and use either the six resulting principal components or the single predictor factor as additional predictor(s)
      end if
    end if
  end if
end if

Continued from previous page

if $Group = true$ **then**
  if $Adaptive = true$ **then**
    if $Control = True$ **then**
      Solve minimization problem (14) for $j = 1, \ldots, 8$ yielding $\hat{Z}_{(1)}, \ldots, \hat{Z}_{(8)}$
      if $Option = 1$ **then**
        Use $\hat{Z}_{(1)}, \ldots, \hat{Z}_{(8)}$to solve minimization problem (15) yielding $\hat{\beta}^{GCALasso1}$
        if $Factor = false$ **then**
          GCALasso1: Use $\sum_{j=1}^{8} \sum_{i=1}^{\hat{N}_j} \hat{\beta}_{i,j}^{GCALasso1} \hat{z}_j^{(i)}$ as a single predictor
        else
          FGCALasso1: Apply PCA as described in Section 3.3.1 to the scaled variables $(\hat{\beta}_{1,1}^{GCALasso1} z_1^{(1)}, \ldots, \hat{\beta}_{\hat{N}_8,8}^{GCALasso1} z_8^{(\hat{N}_8)})$ and use either the six resulting principal components or the single predictor factor as additional predictor(s)
        end if
      else
        Use $\hat{Z}_{(1)}, \ldots, \hat{Z}_{(8)}$to solve minimization problem (15) with $\lambda^{(j)}$ replaced by $\lambda_i^{(j)}$ as specified in Equation (17) yielding $\hat{\beta}^{GCALasso2}$
        if $Factor = false$ **then**
          GCALasso2: Use $\sum_{j=1}^{8} \sum_{i=1}^{\hat{N}_j} \hat{\beta}_{i,j}^{GCALasso2} \hat{z}_j^{(i)}$ as a single predictor
        else
          FGCALasso2: Apply PCA as described in Section 3.3.1 to the scaled variables $(\hat{\beta}_{1,1}^{GCALasso2} z_1^{(1)}, \ldots, \hat{\beta}_{\hat{N}_8,8}^{GCALasso2} z_8^{(\hat{N}_8)})$ and use either the six resulting principal components or the single predictor factor as additional predictor(s)
        end if
      end if
    end if
  end if
end if

Appendix F Bootstrap Procedures

F.1 Bootstrap under the Null Hypothesis

The procedure proposed by Bauer and Hamilton (2018) to robustly test $H_0 : \beta_2 = 0$ in the predictive regression and to compute the size of conventional tests is a parametric bootstrap method that generates data under H_0. The procedure can be divided into six steps which are described below. In the first step, the first $l = 3$ normalized eigenvectors $(w_1, ..., w_l)$ and principal components $(PC_t^{(1)}, ..., PC_t^{(l)})$ corresponding to the variance matrix of the J observed bond yields $i_t = (i_t^{(n_1)}, .., i_t^{(n_J)})'$ are computed. The bond yields can be written in terms of the eigenvectors, principal components and an error term:

$$i_t = W_{J\times l} x_{1t} + v_t \qquad t = 1, ..., T \tag{18}$$

where $W_{J\times l} = (w_1, ..., w_l)$, $W'_{J\times l} W_{J\times l} = I_l$, I_l is the l-dimensional identity matrix, $x_{1t} = (PC_t^{(1)}, ..., PC_t^{(l)})'$, $v_t = (v_t^{(n_1)}, ..., v_t^{(n_J)})'$, and T is equal to the length of the original sample. It follows that the fitted yields are given by $\hat{i}_t = W_{J\times l} x_{1t}$. Secondly, a VAR(1) model is estimated for both x_{1t} and x_{2t} using OLS:

$$x_{it} = \hat{a}_{i0} + \hat{A}_i x_{i,t-1} + e_{it} \qquad i = 1, 2 \text{ and } t = 1, ..., T \tag{19}$$

Thirdly, $N_b = 5,000$ artificial samples of length T are generated. In each bootstrap sample, the bootstrap predictors are constructed as

$$\tilde{x}_{i\tau} = \hat{a}_{i0} + \hat{A}_i \tilde{x}_{i,\tau-1} + \tilde{e}_{i\tau} \qquad i = 1, 2 \text{ and } \tau = 1, ..., T + 12 \tag{20}$$

where $\tilde{x}_{i0} = x_{i0}$ and $(\tilde{e}_{1\tau}, \tilde{e}_{2\tau})$ are randomly drawn with replacement from the joint distribution of (e_{1t}, e_{2t}). The bootstrap yields are constructed in such a way that only the factors in $\tilde{x}_{1\tau}$ have predictive power and the variance structure of the bootstrap yields is similar to that of the observed yields:

$$\tilde{i}_\tau = W_{J\times l} \tilde{x}_{1\tau} + \tilde{v}_\tau \qquad \tau = 1, ..., T + 12 \tag{21}$$

where $\tilde{v}_\tau \overset{iid}{\sim} N(0, \sigma_v^2 I_J)$ and $\sigma_v^2 = \frac{1}{T\cdot J} \sum_{t=1}^{T} \sum_{j=1}^{J} \left(v_t^{(n_j)}\right)^2$. The bootstrap target variable is then generated under the null hypothesis using the bootstrap yields. If the unweighted average excess return across bonds with maturities ranging from 2 to k years is used as the dependent variable

in the application, then the bootstrap dependent variable is constructed as

$$\tilde{y}_{\tau+12}^{H_0} = \frac{1}{k-1}\sum_{n=2}^{k}\widetilde{rx}_{\tau+12}^{(12n)} = \frac{1}{k}\sum_{n=2}^{k}\left(-(12n-12)\tilde{i}_{\tau+12}^{(12n-12)} + 12n\tilde{i}_{\tau}^{(12n)} - 12\tilde{i}_{\tau}^{(12)}\right). \tag{22}$$

Similar to Equation (12), the bootstrap bond excess returns $\widetilde{rx}_{\tau+12}^{(12n)}$ are divided by the number of years to maturity n before being averaged if the weighted average is used as the dependent variable in the application. Fourthly, the predictive regression is estimated in each of the N_b bootstrap samples:

$$\tilde{y}_{\tau+12}^{H_0} = \beta_0 + \beta_1\tilde{x}_{1\tau} + \beta_2\tilde{x}_{2\tau} + u_{\tau+12} \qquad \tau = 1, ..., T \tag{23}$$

Fifthly, the N_b bootstrap samples generated under the null hypothesis are used to test the spanning hypothesis. To this end, the bootstrap p-value and critical value corresponding to a conventional t-test or Wald test for the significance of the coefficients in β_2 are computed. The bootstrap p-value is calculated by

$$\tilde{p} = \frac{1}{N_b}\sum_{j=1}^{N_b}\mathbb{1}_{\{|\tilde{t}^{(j)}|>|t|\}} \tag{24}$$

where the indicator function $\mathbb{1}_{\{x\}}$ is equal to 1 if x is true and 0 otherwise, t denotes the conventional test statistic in the actual data, and $\tilde{t}^{(j)}$ denotes the conventional test statistic in the bootstrap sample j. The bootstrap 5% critical value is given by

$$\tilde{cv} = |\tilde{t}|_{(N_b(1-\alpha))} \qquad 0 \leq \alpha \leq 1 \tag{25}$$

where $\alpha = 0.05$ and $|\tilde{t}|_{(j)}$ denote the order statistic of $|\tilde{t}^{(j)}|$, that is, $|\tilde{t}|_{(1)} \leq |\tilde{t}|_{(2)} \leq ... \leq |\tilde{t}|_{(N_b)}$. Finally, the bootstrap estimate of the size of the conventional test is computed by

$$\tilde{s} = \frac{1}{N_b}\sum_{j=1}^{N_b}\mathbb{1}_{|\tilde{t}^{(j)}|>cv_{5\%}} \tag{26}$$

where $cv_{5\%}$ is the conventional 5% critical value of the test. In the case of a t-test, $cv_{5\%}$ is the 5% critical value corresponding to a Student-t distribution with $T-K$ degrees of freedom, where K is the total number of parameters that are estimated in the predictive regression. In the case of a Wald test, $cv_{5\%}$ is the 5% critical value corresponding to a chi-squared distribution with K_2 degrees of freedom, where K_2 is the number of predictors in x_{2t}.

F.2 Size and Power of Bootstrap Test

In order to calculate the size of the bootstrap test, additional N_b artificial samples are generated from the original N_b bootstrap samples. Specifically, for each bootstrap sample, steps 2-4 described in Section F.1 are repeated using the bootstrapped data instead of the actual data and $N_b = 1$ instead of $N_b = 5,000$. The resulting conventional test statistic in the new artificial sample j is denoted by t_j^*. The size of the bootstrap test is then approximated by

$$s_b = \frac{1}{N_b} \sum_{j=1}^{N_b} \mathbb{1}_{\{|\tilde{t}^{(j)}| > |t^*|_{(N_b(1-\alpha))}\}} \qquad 0 \leq \alpha \leq 1 \tag{27}$$

where $\alpha = 0.05$ and $|t^*|_{(j)}$ denote the order statistic of $|t_j^*|$, that is, $|t^*|_{(1)} \leq |t^*|_{(2)} \leq ... \leq |t^*|_{(N_b)}$.

To compute the power of the bootstrap test, predictive regression (23) is re-estimated in each bootstrap sample using a bootstrap dependent variable that is constructed under the alternative hypothesis:

$$\tilde{y}_{\tau+12}^{H_a} = \tilde{y}_{\tau+12}^{H_0} + \hat{\beta}_2 \tilde{x}_{2\tau}. \tag{28}$$

where $\hat{\beta}_2$ is the estimated coefficient of β_2 in the predictive regression using the actual data. The resulting test statistic of a t-test or Wald test for the significance of the coefficients in β_2 in bootstrap sample j is denoted by $\tilde{t}_{H_a}^{(j)}$. The power of the conventional test and the bootstrap test are now estimated by respectively

$$power_c = \frac{1}{N_b} \sum_{j=1}^{N_b} \mathbb{1}_{|\tilde{t}_{Ha}^{(j)}| > cv_{5\%}} \tag{29}$$

$$power_b = \frac{1}{N_b} \sum_{j=1}^{N_b} \mathbb{1}_{\{|\tilde{t}_{Ha}^{(j)}| > |t^*|_{(N_b(1-\alpha))}\}} \qquad 0 \leq \alpha \leq 1 \tag{30}$$

where $\alpha = 0.05$, and $\mathbb{1}_{\{x\}}$ and $cv_{5\%}$ are defined similarly as in Section F.1.

F.3 Bias-Corrected Bootstrap

In small samples, the autocorrelation of very persistent time series is typically underestimated by OLS (Pope, 1990). As a result, the estimated VAR(1) models in (19) are typically less persistent compared to the true data-generating process. To correct for this bias, the bias-correction method of Kilian (1998) with a bootstrap size of 5,000 is used instead of OLS to estimate these VAR(1) models for x_{1t} and x_{2t}.

Appendix G Simulation Study

G.1 Simulation Study without Overlapping Observations

In the basic setting without overlapping observations, I simulate $N_s = 10,000$ samples of size $T = 100$, estimate the predictive regression $y_{t+1} = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + u_{t+1}$ with $t = 1, ..., T$ in each simulated sample, and study the small-sample properties of the regression coefficients and the test of $H_0 : \beta_2 = 0$. Similar to Bauer and Hamilton (2018), the samples are simulated according to the following data-generating process (DGP):

$$x_{it} = \mu_i + \rho_i x_{i,t-1} + \epsilon_{it} \qquad i = 1, 2 \qquad t = 1, ..., T+1 \tag{31}$$

$$y_t = \epsilon_t^y \qquad t = 2, ..., T+1 \tag{32}$$

where $x_{10} = x_{20} = 0$, $\epsilon_{1t} = \delta v_{3t} + \sqrt{1-\delta^2} v_{1t}$, $\epsilon_{2t} = v_{2t}$, $\epsilon_t^y = v_{3t}$, and $v_{it} \overset{iid}{\sim} N(0,1)$ for $i \in \{1, 2, 3\}$. Hence, in the DGP it holds that $\beta_0 = \beta_1 = \beta_2 = 0$ and $E(\epsilon_{1t}\epsilon_t^y) = E(\epsilon_{1t} v_{3t}) = \delta$. Simulation experiments are conducted for different values μ_1, μ_2, ρ_1, ρ_2, and δ. In each simulation experiment, the correlation between x_{1t} and ϵ_{t+1}^y, the coefficient and standard error bias of $\hat{\beta}_1$ and $\hat{\beta}_2$, the standard deviation of $\hat{\beta}_2$, the size of the test of $H_0 : \beta_2 = 0$ for the conventional OLS t-test and bootstrap test, and the average and standard deviation of the difference in R^2 of the predictive regressions under the alternative hypothesis (R_2^2) and the null hypothesis (R_1^2) are computed. The formulas and procedures used to compute these statistics are given in Section G.2.

G.1.1 No trends

First, I focus on the case in which x_{1t} and x_{2t} do not exhibit a trend. Following Bauer and Hamilton (2018), I use two specifications for x_{it} in this case, namely the AR(1) specification (31) with $\mu_i = 0$ and the the local-to-unity specification proposed by Phillips (1988) and Cavanagh, Elliott and Stock (1995). The local-to-unity specification is used to investigate the effect of the sample length on size distortions and is given by

$$x_{it} = \left(1 + \frac{T(\rho_i - 1)}{T_a}\right) x_{i,t-1} + \epsilon_{it} \qquad i = 1, 2 \tag{33}$$

where T is the finite sample length and T_a is the asymptotic sample length. The local-to-unity asymptotic distribution is obtained by letting $T_a \to \infty$ and is used to approximate the finite sample distribution of the t-statistic for $\hat{\beta}_2$. The reason to use the local-to-unity asymptotic distribution is that its small-sample approximations are substantially better than those corres-

ponding to the conventional first-order asymptotic distribution in the case of near-integrated processes (Chan, 1988). In accordance with Bauer and Hamilton (2018), I use $T_a = 1,000$ because according to Stock (1991) the local-to-unity approximations are accurate even for moderate sample lengths. Bauer and Hamilton (2018) derived that under the null hypothesis the t-statistic for $\hat{\beta}_2$ according to the local-to-unity asymptotic distribution is given by

$$t_a \simeq \delta Z_1 + \sqrt{1-\delta^2} Z_0 \tag{34}$$

where Z_0 and Z_1 are defined and derived in Bauer and Hamilton (2018) .

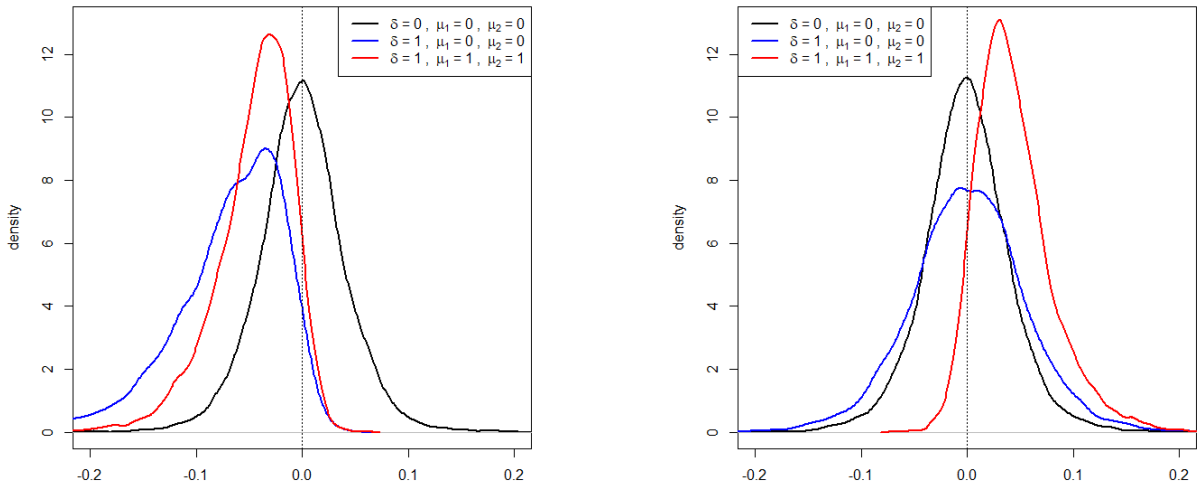
Table G.1
Simulation results for the basic setting without overlapping returns. In this simulation, 10,000 simulation samples of length $T = 100$ are generated according to the data-generating process (DGP) specified in Equations (31) and (32) for different values of δ, ρ_i and μ_i, $i = 1, 2$. In each simulation sample, the predictive regression $y_{t+1} = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + u_{t+1}$ is estimated. The statistics corresponding to the 10,000 regression results are reported in the table. The table reports the correlation between x_{1t} and ϵ^y_{t+1}, the coefficient and standard error bias of $\hat{\beta}_1$ and $\hat{\beta}_2$, the standard deviation of $\hat{\beta}_2$, the size of the test of $H_0 : \beta_2 = 0$ for the conventional OLS t-test and bootstrap test, and the average and standard deviation of the difference in R^2 of the predictive regressions under the alternative hypothesis (R^2_2) and the null hypothesis (R^2_1) are computed. The formulas and procedures used to compute these statistics are given in Appendix G.2.

			Corr.	Coefficient Bias		SE bias (%)		Std.	Size			$R^2_2 - R^2_1$	
ρ_1	ρ_2	δ	$(x_{1t}, \epsilon^y_{t+1})$	β_1	β_2	β_1	β_2	β_2	Sim.	Asym.	Boot.	Mean	Std.
$\mu_1 = 0$, $\mu_2 = 0$													
0.99	0.99	0.0	-0.001	-0.000	-0.001	-5.5	-5.5	0.041	0.049	0.047	0.047	0.01	0.01
0.00	0.00	1.0	-0.012	-0.012	-0.002	1.9	-1.5	0.103	0.052	0.049	0.048	0.01	0.01
0.99	0.00	1.0	-0.107	-0.042	-0.001	-18.8	-0.9	0.102	0.050	0.053	0.047	0.01	0.01
0.99	0.80	1.0	-0.107	-0.045	0.001	-20.8	-8.1	0.073	0.066	0.068	0.061	0.01	0.02
0.90	0.90	1.0	-0.062	-0.052	-0.000	-14.3	-15.3	0.063	0.086	0.090	0.058	0.01	0.02
0.99	0.99	0.8	-0.108	-0.055	-0.001	-21.5	-22.9	0.050	0.109	0.113	0.068	0.01	0.02
0.99	0.99	1.0	-0.134	-0.068	0.001	-27.7	-29.4	0.054	0.149	0.149	0.080	0.02	0.02
$\mu_1 = 0$, $\mu_2 = 1$													
0.99	0.99	0.0	-0.001	-0.001	0.000	-4.6	-5.9	0.008	0.052		0.050	0.01	0.01
0.00	0.00	1.0	-0.009	-0.009	0.000	2.0	-1.1	0.103	0.050		0.050	0.01	0.01
0.99	0.00	1.0	-0.108	-0.043	0.000	-18.9	0.4	0.101	0.048		0.049	0.01	0.01
0.99	0.80	1.0	-0.108	-0.046	-0.000	-20.7	-8.9	0.068	0.071		0.056	0.01	0.02
0.90	0.90	1.0	-0.061	-0.052	-0.000	-13.5	-16.6	0.047	0.090		0.054	0.01	0.02
0.99	0.99	0.8	-0.108	-0.070	0.000	-17.7	-41.7	0.012	0.180		0.077	0.02	0.02
0.99	0.99	1.0	-0.132	-0.087	-0.000	-23.7	-50.5	0.015	0.267		0.080	0.03	0.03
$\mu_1 = 1$, $\mu_2 = 0$													
0.99	0.99	0.0	-0.001	0.000	0.001	-4.3	-4.5	0.046	0.052		0.048	0.01	0.01
0.00	0.00	1.0	-0.011	-0.011	-0.001	0.9	-0.2	0.102	0.051		0.045	0.01	0.01
0.99	0.00	1.0	-0.019	-0.001	0.001	-3.0	0.6	0.101	0.048		0.048	0.01	0.01
0.99	0.80	1.0	-0.018	-0.001	0.001	-2.5	-1.3	0.069	0.049		0.052	0.01	0.01
0.90	0.90	1.0	-0.047	-0.029	0.001	-10.3	-10.6	0.060	0.072		0.058	0.01	0.02
0.99	0.99	0.8	-0.018	-0.002	-0.001	-6.4	-5.9	0.046	0.049		0.047	0.01	0.01
0.99	0.99	1.0	-0.024	-0.003	0.000	-9.2	-7.1	0.047	0.056		0.048	0.01	0.01
$\mu_1 = 1$, $\mu_2 = 1$													
0.99	0.99	0.0	-0.002	-0.000	0.000	-3.5	-3.3	0.032	0.048		0.047	0.01	0.01
0.00	0.00	1.0	-0.010	-0.010	0.001	-0.1	-1.8	0.103	0.055		0.056	0.01	0.01
0.99	0.00	1.0	-0.020	-0.001	0.001	-1.6	-1.5	0.103	0.054		0.052	0.01	0.01
0.99	0.80	1.0	-0.020	-0.001	0.004	-2.4	-1.1	0.065	0.050		0.050	0.01	0.01
0.90	0.90	1.0	-0.049	-0.038	0.017	-12.3	-12.5	0.050	0.082		0.051	0.01	0.02
0.99	0.99	0.8	-0.019	-0.035	0.035	-12.0	-11.7	0.035	0.162		0.053	0.02	0.02
0.99	0.99	1.0	-0.023	-0.045	0.043	-17.0	-15.8	0.036	0.238		0.058	0.03	0.03

I also simulate $N_s = 10,000$ samples of size T_a based on specification (33) and estimate the corresponding asymptotic t-statistic given in Equation (34) using Rieman sums to approximate the integrals in Z_0 and Z_1. The simulation estimate of the size for the conventional t-test according to the local-to-unity asymptotic distribution is computed using Equation (35) where

$t^{(q)}$ is replaced by $t_a^{(q)}$, which is the value for t_a in qth simulated sample.

In the top panel of Table G.1 the simulation results in case of no trends are given for different values of ρ_i and δ. From these results, it follows that the true size of the t-test of $\beta_2 = 0$ according to the small-sample simulations is very close to the nominal size of 5% if x_{1t} and x_{2t} are either not autocorrelated ($\rho_i = 0$) or strictly exogenous ($\delta = 0$). Even in case x_{1t} is highly persistent ($\rho_1 = 0.99$), the true size is still equal to 5% if x_{2t} is not serially correlated ($\rho_2 = 0$). However, if $\rho_i \neq 0$ and $\delta \neq 0$ the true size is larger than 5% and increases in ρ_i and δ. For instance, the true size equals approximately 15% if $\rho_i = 0.99$ and $\delta = 1$, which means that in this case H_0 is rejected at a frequency that is about three times larger than it should be.



(a) $\hat{\beta}_1$ **(b)** $\hat{\beta}_2$

Figure G.1
Simulation distibution of $\hat{\beta}_1$ (on the left) and $\hat{\beta}_2$ (on the right) for three different scenarios. In this simulation, 10,000 simulation samples of length $T = 100$ are generated according to the data-generating process (DGP) specified in Equations (31) and (32) with $\rho_1 = \rho_2 = 0.99$ and for different values of δ, μ_1 and μ_2. In each simulation sample, the predictive regression $y_{t+1} = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + u_{t+1}$ is estimated and the resulting $\hat{\beta}_1$ and $\hat{\beta}_2$ are included in the density plots.

A similar narrative holds for the coefficient bias in $\hat{\beta}_1$. If the regressors are either not autocorrelated or strictly exogenous, then there is no bias in $\hat{\beta}_1$. Otherwise, $\hat{\beta}_1$ is biased and the bias increases in ρ_i and δ. In contrast, $\hat{\beta}_2$ remains unbiased. This is also illustrated by Figure G.1 which plots the simulation distribution of $\hat{\beta}_1$ and $\hat{\beta}_2$ in case of $\mu_i = 0$ and $\delta = 1$ for both $\rho_i = 0$ and $\rho_i = 0.99$. The figure shows that both $\hat{\beta}_1$ and $\hat{\beta}_2$ are unbiased if $\rho_i = 0$, whereas only $\hat{\beta}_2$ is unbiased if $\rho_i = 0.99$. Therefore, size distortions in the test of $\beta_2 = 0$ are not caused by the Stambaugh (1999) coefficient bias, and there must be a different source for these distortions. Panel A in Table G.1 shows that the standard error bias in $\hat{\beta}_2$ has a similar pattern as the size distortions: it is small if either ρ_i or δ is equal to zero, and it increases in ρ_i and δ. For example, if $\rho_i = 0.99$ and $\delta = 1$ the average bias in the simulation is equal to about -30%, meaning that the estimates of the OLS standard errors are on average 30% lower than the standard deviation

of the N_s estimated coefficients in the simulation experiment. Thus, the size distortions are caused by a downward bias in the standard errors and not by a coefficient bias.

The size distortions are decreased when using the bootstrap test. The size of the bootstrap test is namely quite close to or slightly above 5% in all scenarios. In the worst-case scenario with $\rho_i = 0.99$ and $\delta = 0$, the size of the bootstrap test equals 8.0% compared to a true size according to the small-sample simulations of 14.9%. In this case, the regressors are very persistent and the bias-corrected bootstrap method is useful. In fact, using the bias-correction procedure of Kilian (1998) with only 25 bootstrap replications for each sample can already reduce the bootstrap test size to 6.2% in this scenario.

The first panel of Table G.1 shows that the estimates of the size according to the local-to-unity distribution are very similar to the small-sample simulations. Following Bauer and Hamilton (2018), the true size of the t-test according to the local-to-unity distribution is plotted for sample lengths from 50 to 1000. The results for $\delta = 1$ and three different values of $\rho = \rho_1 = \rho_2$ are given in Figure G.2. If $\rho = 1$, the true size stays constant at around 17% and is not impacted by the sample length. In contrast, if $\rho \leq 0$ the true size decreases towards 5% when the sample length increases and this convergence is faster for smaller values of ρ. Hence, the size distortions decrease with sample size if the regressors do not possess a unit root.

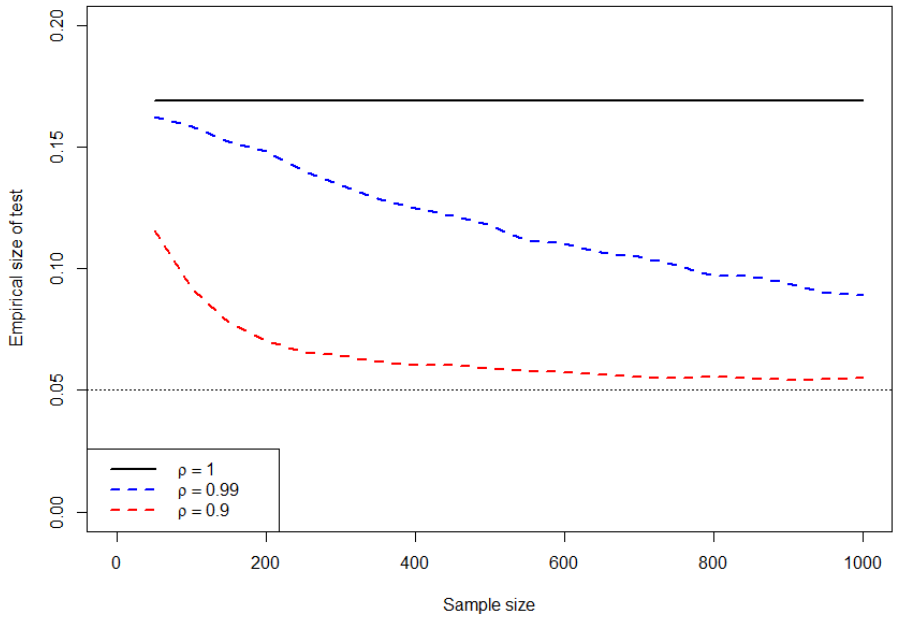


Figure G.2
Simulation estimate of the size of the conventional t-test of $H_0 : \beta_2 = 0$ corresponding to the local-to-unity asymptotic distribution in the basic setting without overlapping returns for different sample sizes. To calculate this estimate, 100,000 simulation samples of length $T = 100$ are generated according to the data-generating process (DGP) specified in Equations (33) and (32) with $\delta = 1$, $c_i = T(\rho_i - 1)$ and $\rho = \rho_1 = \rho_2$. The nominal size is equal to 5%.

G.1.2 Trends

In this section, the small sample effects of trends in the regressors on the size distortions are explored. If $\mu_i \neq 0$ and $\rho_i \leq 1$, x_{it} is a stationary AR(1) process with the tendency to revert to its unconditional mean equal to $\mu_i/(1-\rho_i)$. Since the process is initialized at $x_{i0} = 0$ in the simulations, a trend is induced in the small samples. For example, if $\mu_i = 1$ and $\rho_i = 0.99$ the unconditional mean is equal to 100 and the process tends to rise from 0 to 100. In such a setting, x_{it} is dominated by a deterministic time trend.

The second panel in Table G.1 shows that the Stambaugh bias in $\hat{\beta}_1$ and the standard error bias in $\hat{\beta}_2$ are exacerbated if $\mu_1 = 0$ and $\mu_2 = 1$. The larger standard error bias in $\hat{\beta}_2$ leads to larger size distortions compared to the case in which $\mu_2 = 0$. In fact, the true size according to the small-sample simulations nearly doubles from 14.9% to 26.7% when $\rho_i = 0.99$ and $\delta = 1$. In contrast, the size of the bootstrap test seems to be unaffected by the trend in x_{2t}. Again, the bias-corrected bootstrap is able to decrease the size distortion in the presence of persistent regressors: its size is equal to 5.6% compared to 8.0% for the simple bootstrap test when $\rho_i = 0.99$ and $\delta = 1$. Since the correlation between x_{1t} and ϵ^y_{t+1} is not affected by the trend in x_{2t}, the larger size distortions in the conventional test can only be explained by the presence of the trend in x_{2t}.

In contrast, the size distortions nearly disappear if only x_{1t} exhibits a trend. In this case, the correlation between x_{1t} and ϵ^y_{t+1} is relatively close to zero (see Panel C), because the random error term in the AR(1) process of x_{1t} is dominated a the time trend if ρ_1 is close to one. As a result, the issue of no strict exogeneity underlying the predictive regression is less severe. However, conclusions change if x_{2t} is also trending. Bauer and Hamilton (2018) show that the distribution of $\hat{\beta}_1$ is the same as the distribution of negative $\hat{\beta}_2$ if $\mu_1 = \mu_2 = 1$. This is illustrated in Figure G.1. Moreover, this is shown by the mirror images of the coefficient biases and the roughly equal standard error biases in $\hat{\beta}_1$ and $\hat{\beta}_2$ across all cases in Panel D of Table G.1.

From this panel, it also follows that a trend in both x_{1t} and x_{2t} leads to size distortions that are approximately equal to the size distortions reported in Panel B, even though the standard error bias in $\hat{\beta}_2$ is substantially lower. The similar size distortions can be explained by the fact that in this case there is not only a bias in the standard errors of $\hat{\beta}_2$ but also a bias in the coefficient $\hat{\beta}_2$ itself. This is in accordance with Bauer and Hamilton (2018), who show that including two trending regressors in the predictive regression leads to spurious results and distorts the conventional inference. Ultimately, it can be concluded from the last two panels that the bootstrap test is fairly reliable when x_{1t} exhibits a trend, as all estimates of its size fall within the range of 4.5% and 6.0%.

G.2 Statistics for Simulation Study

The estimated coefficient bias is computed as $\frac{1}{N_s}\sum_{q=1}^{N_s}(\hat{\beta}_i^{(q)} - \beta_i)$ where $\hat{\beta}_i^{(q)}$ is the estimated coefficient for β_i in the predictive regression in the qth simulated sample. The standard error bias is estimated by $\frac{1}{N_s}\sum_{q=1}^{N_s}\left(\frac{se_i^{(q)}}{\hat{\sigma}(\hat{\beta}_i)}\right) - 1$ where $se_i^{(q)}$ is the estimated conventional OLS standard error of $\hat{\beta}_i^{(q)}$ and $\hat{\sigma}(\hat{\beta}_i)$ is the standard deviation of the coefficient estimates for β_i across the N_s simulated samples. Similar to the bootstrap estimate of the size of a test specified in Equation (26), the simulation estimate of the size for the conventional OLS t-test of $H_0 : \beta_2 = 0$ is given by:

$$\hat{s} = \frac{1}{N_s}\sum_{q=1}^{N_s} \mathbb{1}_{|t^{(q)}|>t\left(1-\frac{\alpha}{2},\, T-3\right)} \tag{35}$$

where $t^{(q)}$ is the conventional OLS t-statistic in qth simulated sample, and $t(1-\alpha/2,\ T-3)$ is the critical value corresponding to a Student-t distribution with $T-3$ degrees of freedom and significance level α. Finally, to calculate the size of the bootstrap test in the simulation experiments, a modified version of the procedure described in Section F.2 is used. In this adjusted procedure, described in the next subsection, h is set equal to one in the setting of no overlapping returns and equal to twelve in the setting of overlapping returns.

G.2.1 Procedure to Calculate the Size of the Bootstrap Test in the Simulation Experiments

In order to calculate the size of the bootstrap test in the simulation study, additional N_s artificial samples are generated from the original N_s simulated samples generated under the null hypothesis. Specifically, the following steps are conducted for each simulated sample j. First, the following models for x_{1t}, x_{2t} and y_t are fitted using OLS:

$$\begin{aligned} x_{it} &= \hat{\mu}_i + \hat{\rho}_i x_{i,t-1} + e_{it} && i = 1,2 \quad t = 1,...,T \\ y_{t+h} &= \hat{\beta}_0 + \hat{\beta}_1 x_{1t} + \hat{u}_{t+h} && t = 1,...,T \end{aligned}$$

Second, the estimated parameters are used to generate one new artificial sample of length T:

$$\tilde{x}_{i\tau} = \hat{\mu}_i + \hat{\rho}_i \tilde{x}_{i,\tau-1} + \tilde{e}_{i,\tau} \qquad i = 1,2 \quad \tau = 1,...,T$$

$$\tilde{y}_{\tau+h} = \begin{cases} \hat{\rho}_1 \tilde{x}_{1\tau} + \tilde{u}_{\tau+1} & \text{if } h = 1 \\ \tilde{x}_{1\tau} + \frac{1}{2}\sum_{s=\tau-h+1}^{\tau-1} \tilde{u}_s + \tilde{u}_{\tau+h} & \text{if } h \geq 2 \end{cases} \qquad \tau = 1,...,T$$

where $\tilde{x}_{i0} = 0$, $(\tilde{e}_{1,\tau}, \tilde{e}_{2,\tau}, \tilde{u}_\tau)$ are randomly drawn with replacement from the joint distribution of $((e_{1,t}, e_{2,t}, \hat{u}_t)$. As a result, the contemporaneous correlation between the residuals in the simulated data is maintained. Third, the predictive regression on this new artificial sample $\tilde{y}_{\tau+h} = \beta_0 + \beta_1 \tilde{x}_{1\tau} + \beta_2 \tilde{x}_{2\tau} + u_{\tau+h}$ with $\tau = 1, ..., T$ is estimated and the t-statistic t_q^* corresponding to a conventional test for the significance of parameter β_2 is computed. The t-statistic is computed using OLS standard errors if $h = 1$ and using Newey-West standard errors if $h \geq 2$.

After repeating the procedure described above for $j = 1, ..., N_s$, the size of the bootstrap test in the simulation study can be estimated by

$$s_b = \frac{1}{N_s} \sum_{q=1}^{N_s} \mathbb{1}_{\{|t^{(q)}| > |t_q^*|_{(N_s(1-\alpha))}\}} \qquad 0 \leq \alpha \leq 1 \tag{36}$$

where $\alpha = 0.05$, $t^{(q)}$ is the conventional t-statistic in qth simulated sample, and $|t^*|_{(j)}$ denote the order statistic of $|t_j^*|$, that is, $|t^*|_{(1)} \leq |t^*|_{(2)} \leq ... \leq |t^*|_{(N_s)}$.

Appendix H Preliminary Empirical Analysis

Table H.1 reports some preliminary statistics for the four published studies and the most relevant factor extraction methods and indicates for each application over which sample period the statistics are computed. Firstly, the first-order autocorrelations of x_{1t} and x_{2t} are computed. These statistics are presented in the second and third columns of the table. The first and second yield curve PCs are very persistent in all applications with first-order autocorrelations of around 0.98 and 0.95 respectively. The persistence of the additional predictors varies across applications. In the case of JPS and CPO, the additional predictors are also very persistent with autocorrelations of 0.986 and 0.998 for respectively INF and τ. This substantiates the decision to apply the bias-corrected bootstrap approach in these applications. In the case of CP, the persistence is substantially lower with first-order autocorrelations of 0.425 and 0.227 for the fourth and fifth yield PC respectively. For the other applications, the persistence of the additional predictors is also lower compared to JPS and CPO, but it is still considerable.

Table H.1
First-order autocorrelations of the independent variables in the predictive regression, the average of the 11 estimated coefficients corresponding to a MA(11) model for the fitted residuals of the predictive regression, and the number of observations for different applications over their earlier samples. The statistics are computed using the original data of the published studies for the first four applications and using the new extended large macroeconomic data set for the other applications. If x_{2t} contains more than two variables, which is indicated with the asterisk *, the two autocorrelations with the largest values in absolute terms are reported.

	First-order Autocorrelations		MA(11) coefficients	
	$(PC_t^{(1)}, PC_t^{(2)}, PC_t^{(3)})$	x_{2t}	Average	#obs.
Original Sample				
JPS	(0.974, 0.976, 0.815)	(0.91, 0.986)	0.625	276
CPO	(0.987, 0.944, 0.773)	(0.998)	0.457	470
CP	(0.98, 0.94, 0.592)	(0.425, 0.227)	0.401	468
LN	(0.984, 0.944, 0.600)	(0.766, 0.748)*	0.493	516
1964-2007				
PCA	(0.984, 0.943, 0.574)	(0.816)	0.570	516
sPCA	(0.984, 0.943, 0.574)	(0.92)	0.625	516
FLasso	(0.984, 0.943, 0.574)	(0.774)	0.582	516
GCALasso1	(0.984, 0.943, 0.574)	(0.833)	0.295	516
GCALasso2	(0.984, 0.943, 0.574)	(0.846)	0.285	516

As argued in Section 3.5 there are endogeneity issues in all applications, as the yield curve PCs violate the strict exogeneity condition by construction. Additionally, using annual overlapping average excess bond return as dependent variable in the predictive regression leads to another econometric problem. This is illustrated in the fourth column of Table H.1 which provides the average of the 11 coefficient estimates of the MA(11) model for the residuals resulting from the predictive regression. In accordance with theory, the average value is non-zero in all applications indicating that the error terms are indeed correlated. Furthermore, given that the average value of this column is around 0.5, this demonstrates that the choice to generate the residuals ϵ_t^y according to a MA(11) model with coefficient values equal to 0.5 in the simulations with overlapping observations is reasonable. Lastly, the fifth column gives the number of obser-

vations in the considered samples and shows that especially the original sample corresponding to JPS is relatively small.

Due to the presence of strong trends in the CPO application, I pay special attention to this application. Figure 13 plots both the yield on a 10-year zero-coupon bond and the measure of trend inflation used in CPO. It shows that the time series exhibit an upward trend until the early 1980s and a downward trend thereafter. Fitting an AR(1) model for the trend inflation over the 1985-2016 sample results in $\hat{\rho}_2 = 0.99$ and $\hat{\mu}_2/\hat{\sigma}_2 = 1.8$. The drift relative to the standard deviation of the error term is thus stronger than the value $\mu_2/\sigma_2 = 1$ used in the simulation study. Moreover, the value of the trend inflation in January 1985 is about 5 times larger than $\hat{\mu}_2/(1-\hat{\rho}_2)$ which implies that the downward drift over 1985-2013 is approximately 4 times larger in this application than in the simulation study. Therefore, the size distortions are expected to be relatively large in the CPO application, as the simulation study demonstrated that even in case of a weaker trend, the size distortions are already substantial.

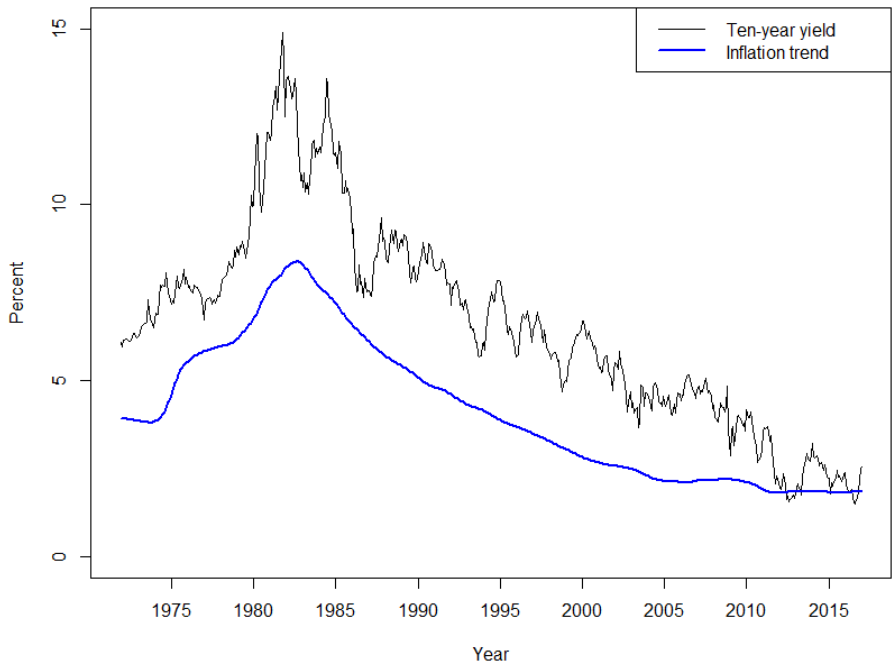


Figure H.1
10-year bond yield along with the inflation trend as specified by Equation 13

Appendix I Empirical Analysis for JPS, CPO and CP

I.1 In-Sample Analysis

The in-sample coefficients and statistics according to both the conventional test and bootstrap test of the additional predictors x_{2t} in the predictive regression corresponding to JPS, CPO and CP are given in Table I.1. These results are used to formally test the spanning hypothesis. Some results may slightly differ from the results obtained by Bauer and Hamilton (2018) since the random seeds used as input for the bootstrap procedures are not always the same.

Table I.1
The in-sample coefficients and statistics of the additional predictors x_{2t} in the predictive regressions corresponding to JPS, CPO and CP estimated over the original sample and the 1985-2016 sample. The variables used in the predictive regression for each application are given in Table C.1. The data sets provided by Bauer and Hamilton (2018) are used to estimate the regressions. Under *Wald* it reports the statistics corresponding to the test of joint significance of these additional predictors in case the number of additional predictors in an application is larger than one. The conventional statistics and p-values are computed using Newey-West standard errors with 18 lags, except for CPO. In CPO, the reverse-regression approach of Wei and Wright (2011) is used to compute the standard errors. The bias-corrected bootstrap is used in JPS and CPO, while the simple bootstrap is employed in CP. The conventional size and power are estimated using Equation (26) and (29) respectively. The bootstrap 5% critical values (c.v.'s) and p-values are computed using Equation (25) and (24). The size and power of the bootstrap test are approximated using Equations (27) and (30).

			Test Stat.	5% c.v.	p-value		Size		Power	
		Coefficient	Conv.	Boot.	Conv.	Boot.	Conv.	Boot.	Conv.	Boot.
JPS										
1985-2008	GRO	-2.200	-2.475	3.263	0.014	0.123	0.217	0.061	0.446	0.201
	INF	-6.052	-4.265	4.161	0.000	0.045	0.318	0.070	0.979	0.887
	Wald		25.152	25.364	0.000	0.052	0.411	0.066	0.986	0.875
1985-2016	GRO	-0.429	-0.537	3.137	0.592	0.714	0.196	0.062	0.221	0.066
	INF	-2.420	-1.798	3.729	0.073	0.327	0.278	0.067	0.786	0.515
	Wald		3.350	21.460	0.187	0.555	0.361	0.069	0.820	0.479
CPO										
1971-2011	τ	-0.962	-6.329	3.572	0.000	0.000	0.423	0.078	0.998	0.979
1985-2016	τ	-0.607	-3.708	3.640	0.000	0.043	0.405	0.088	0.988	0.909
CP										
1964-2003	$PC^{(4)}$	-16.128	-5.626	2.214	0.000	0.000	0.086	0.049	0.996	0.989
	$PC^{(5)}$	-2.038	-0.748	2.194	0.455	0.511	0.081	0.047	0.147	0.104
	Wald		31.919	7.920	0.000	0.000	0.103	0.048	0.993	0.983
1985-2016	$PC^{(4)}$	-9.585	-1.460	2.397	0.145	0.221	0.106	0.045	0.484	0.317
	$PC^{(5)}$	-9.360	-1.263	2.379	0.207	0.299	0.104	0.047	0.418	0.286
	Wald		4.180	9.592	0.124	0.258	0.142	0.047	0.608	0.391

From the top panel of this table, it follows that according to the conventional tests the additional predictors x_{2t} in JPS are both individually and jointly significant over the 1985-2008 sample. Notably, the p-value of the Wald test, which tests for joint significance of the additional predictors, is lower than 0.1%. However, as indicated in Section 3.5 and illustrated in Appendix H, some econometric problems distort the size of these conventional tests. To determine the magnitude of these size distortions and to robustly test the spanning hypothesis within this application, the bias-corrected bootstrap is utilized. The bootstrap results expose that the true sizes of the conventional tests using the Newey-West standard errors with 18 lags are equal to 22-41% instead of the nominal size of 5%. In contrast, the estimated sizes of the bootstrap tests are slightly above 5%. Furthermore, the bootstrap tests result in much weaker evidence against the spanning hypothesis than the conventional tests. Namely, the bootstrap p-values indicate

that the estimated coefficient on GRO lacks significance, even at 10% level, and the coefficient on INF exhibits marginal significance at 5% level. Furthermore, the p-value corresponding to the bootstrap Wald test is slightly above 5%. Bauer and Hamilton (2018) also consider the simple bootstrap in their analysis for JPS. The results for the simple bootstrap are similar to those for the bias-corrected bootstrap.

The top panel of Table I.2 shows that the adjusted R^2 increases from 19% to 38% in the 1985-2008 sample when including GRO and INF to the regression with the three yield PCs as independent variables. Even though this increase of 19 percentage points is quite substantial, it is below the upper bound of the bootstrap interval for $R_2^2 - R_1^2$ for both the simple and bias-corrected bootstrap and therefore it does not provide significant evidence against the spanning hypothesis at 5% level. The fact that even increases of 19 percentage points are not uncommon under the null hypothesis also illustrates the large variability of the adjusted R^2 in this application. Looking at the 1985-2016 sample, it follows that the increase in the adjusted R^2 when including x_{2t} is far from being significant and considerably smaller than in the original JPS sample. The weakening of the evidence against the spanning hypothesis in the later sample is also visible in Table I.1. In this sample, neither the conventional tests nor the bootstrap tests result in significant results at 5% level.

Table I.2
In-sample adjusted R^2 for the restricted regression model with only x_{1t} (R_1^2), the adjusted R^2 for the unrestricted regression model including both x_{1t} and x_{2t} (R_2^2), and the difference in adjusted R^2 ($R_2^2 - R_1^2$) corresponding to JPS, CPO and CP estimated over the original sample and the 1985-2016 sample. The variables used in the predictive regression for each application are given in Table C.1. The data sets provided by Bauer and Hamilton (2018) are used to estimate the regressions. The left half of the table provides the results for the earlier sample periods and the right half of the table provides the results for the later sample periods. For each application, the first row reports the adjusted R^2 statistic in the corresponding actual data set; the second and third rows report respectively the mean and 95%-quantiles of the statistics in the 5,000 bootstrap replications under H_0.

	R_1^2	R_2^2	$R_1^2 - R_2^2$	R_1^2	R_2^2	$R_1^2 - R_2^2$
JPS	Original sample, 1985-2008			Later sample, 1985-2016		
Data	0.19	0.38	0.19	0.17	0.21	0.04
Bootstrap	0.32	0.38	0.06	0.28	0.33	0.05
	(0.10, 0.55)	(0.15, 0.60)	(-0.00, 0.20)	(0.08, 0.49)	(0.12, 0.53)	(-0.00, 0.17)
BC bootstrap	0.36	0.42	0.06	0.29	0.34	0.05
	(0.09, 0.64)	(0.14, 0.67)	(-0.00, 0.22)	(0.06, 0.53)	(0.11, 0.57)	(-0.00, 0.20)
CPO	Original sample, 1971-2007			Later sample, 1985-2016		
Data	0.16	0.50	0.33	0.17	0.34	0.17
Bootstrap	0.18	0.25	0.07	0.28	0.34	0.06
	(0.03, 0.39)	(0.08, 0.44)	(-0.00, 0.21)	(0.06, 0.52)	(0.12, 0.56)	(-0.00, 0.23)
CP	Original sample, 1964-2003			Later sample, 1985-2016		
Data	0.26	0.35	0.09	0.15	0.18	0.03
BC Bootstrap	0.21	0.22	0.01	0.30	0.31	0.01
	(0.06, 0.40)	(0.06, 0.41)	(0.00, 0.02)	(0.09, 0.52)	(0.11, 0.52)	(0.00, 0.05)

Next, I discuss the in-sample results regarding CPO. In this application, the reverse-regression (RR) delta method of Wei and Wright (2011) is employed to alleviate the econometric problem related to overlapping returns. Furthermore, the bias-corrected bootstrap is utilized, because in Appendix H it is established that x_{2t} is very persistent in this application. From the second

panel of Table I.1 it can be seen that τ_t is highly significant according to both the conventional and bootstrap test in the 1971-2011 sample period. Even though the conclusions of the conventional and bootstrap tests are the same, their properties differ substantially. The estimate of the true size for the conventional test using the reverse regression approach is approximately equal to 42% whereas the size of the bootstrap test is much closer to 5%.

Table I.3
The in-sample coefficients and statistics of the additional predictors x_{2t} in the alternative predictive regressions corresponding to CPO estimated over the original sample starting in January 1971 and ending in December 2011. In these predictive regressions, the dependent variable is $\overline{wrx}^{(}_{t+h}15)$, x_{1t} consist of a constant and the first three PCs of the yields with 1 month and 1 to 15 years maturity, and $x_{2t} = \tau_t$. The regressions are estimated using the data set employed by Bauer and Hamilton (2018) to revisit the evidence of Cieslak and Povala (2015). Due to the high persistence of τ_2, the bias-corrected bootstrap is used. The table reports the coefficients and statistics for τ in different scenarios. The second column indicates whether the bootstrap samples are initialized at the first values of the actual sample or the population means implied by the coefficients of the VAR(1) model estimated from the full sample. Using the notation of Appendix F, it holds that in the first case $\hat{x}_{i0} = x_{i0}$ and in the second case $\hat{x}_{i0} = \hat{a}_{i0}(I_{K_i} - \hat{A}_i)^{-1}$ where K_i is the number of variables contained in x_{it}. The third column specifies the value for h. In case $h = 1$, $\overline{wrx}^{(15)}_{t+1}$ is computed using the 1-month Treasury bill interest rate and the approximation $i_{n-1,t+1} \approx i_{n,t+1}$. The fourth column indicates which standard errors are used: *NW* stands for Newey-West standard errors with 18 lags, *RR* stands for Reverse-Regression standard errors, and *Wh* stands for White standard errors. The conventional size and power are estimated using Equation (26) and (29) respectively. The bootstrap 5% critical values (c.v.'s) and p-values are computed using Equation (25) and (24). The size and power of the bootstrap test are approximated using Equations (27) and (30).

					Stat.	5% c.v.	p-value (in %)		Size (in %)		Power (in %)	
	$\tilde{x}_{i0}$	h	AC	Coef.	Conv.	Boot.	Conv.	Boot.	Conv.	Boot.	Conv.	Boot.
CPO												
τ	x_{i0}	12	NW	-0.962	-7.664	4.507	0.000	0.001	0.535	0.078	1.000	0.979
τ	x_{i0}	1	Wh	-0.104	-4.063	3.090	0.000	0.002	0.330	0.086	0.327	0.084
τ	$\hat{a}_{i0}(I_{K_i} - \hat{A}_i)^{-1}$	12	RR	-0.962	-6.329	2.759	0.000	0.000	0.164	0.078	0.308	0.203
τ	$\hat{a}_{i0}(I_{K_i} - \hat{A}_i)^{-1}$	1	Wh	-0.104	-4.063	2.501	0.000	0.001	0.129	0.082	0.124	0.080

To investigate the sources of the enormous size distortion in the conventional test, I closely follow Bauer and Hamilton (2018) and conduct some additional investigation. The results are provided in Table I.3. Firstly, the magnitude of the problem related to overlapping returns is examined. Using the Newey-West standard errors with 18 lags instead of the RR standard errors leads to an even larger true size of about 54%. It thus follows that the problem related to the overlapping returns can be mitigated by using RR standard errors. However, the RR standard errors do not completely solve the issue, as in case of no overlapping returns the estimate of the true size decreases to 33%. The latter result is obtained by setting $h = 1$ and computing y_{t+1} instead of y_{t+12} using the 1-month Treasury bill interest rate and the approximation $i_{n-1,t+1} \approx i_{n,t+1}$. In this case, White standard errors are used in the t-test. Secondly, the impact of the presence of trends in the predictors is assessed. Removing the trends in the predictors by initializing the bootstrap samples at the population means according to the VAR(1) model estimated over the complete sample results in a true size of 16.4%. In the absence of both overlapping returns and trends, the estimate of the true size reduces even further to 12.9%. These results show that the enormous size distortions are primarily due to the presence of trends.

From Table I.1 it follows that in CPO the significance of x_{2t} is also less strong in the 1985-

2016 sample. In particular, the bootstrap p-value corresponding to the estimated coefficient on τ_t is equal to 4.3% and is thus only marginally significant at 5% level. Furthermore, the second panel of Table I.2 shows that the predictive power of τ_t beyond the three yield PCs is weaker in the later sample. The increase in the adjusted R^2 is namely smaller and becomes insignificant in the later sample.

Moving forward, I delve into the analysis for CP. The third panel of Table I.1 shows that the size distortions are relatively small in this application. This can be explained by the low persistence of the fourth and fifth yield PCs. In the original sample, the fourth PC is highly significant and the fifth PC is far from being significant according to both the conventional test and the bootstrap test. Furthermore, the joint statistical significance of the fourth and fifth PCs as measured by the Wald test is strong. However, all the significant results in this sample are insignificant in the later sample. The third panel of Table I.2 also reveals that the additional predictive power of the fourth and fifth yield PCs is only significant in the earlier sample.

I.2 Out-of-Sample Analysis

The out-of-sample results corresponding to JPS, CPO and CP are given in Table I.4. As discussed in Section 3.6, Bauer and Hamilton (2018) use the three yield PCs that are estimated over the full sample, including the out-of-sample period, to calculate out-of-sample forecasts. This leads to a look-ahead bias. To correct for this look-ahead bias, I recursively estimate the three yield PCs based on information that is available in month t to calculate the out-of-sample forecast for the excess bond return in month $t + 12$. Consequently, the results in the table slightly deviate from the out-of-sample results obtained by Bauer and Hamilton (2018). The table shows that even though including the additional predictors x_{2t} in the original in-sample predictive regressions improves the mean squared error (MSE), it deteriorates the out-of-sample MSE.. The deterioration of the prediction error is significant at 10% level according to the DM test in the three applications and even strongly significant in the case of JPS and CPO.

Table I.4
In-sample and out-of-sample performance of the predictive regressions corresponding to JPS, CPO and CP. The variables used in the predictive regression for each application are given in Table C.1. The in-sample period is the original sample period that is employed in each study. The out-of-sample period starts one month later than the end of the in-sample period and ends in December 2016. To generate the out-of-sample forecasts, expanding window estimation is used. The table reports the start of the in-sample and out-of-sample period in each application. It also reports the adjusted R^2 for the restricted model with the three yield PCs as predictors (R_1^2) and the adjusted R^2 for the unrestricted model including both the three yield PCs and the additional predictors (R_2^2). Under *MSE ratio* and *p-value* the table reports respectively the mean-squared errors for the unrestricted model relative to the mean-squared errors for the restricted model and the p-values of the Diebold-Mariano test for equal prediction accuracy of the two models.

	In-sample				Out-of-sample		
	Start	R_1^2	R_2^2	MSE ratio	Start	MSE ratio	DM p-value
JPS	198501	0.189	0.380	0.758	200801	2.213	0.005
CPO	197111	0.165	0.495	0.603	201101	3.285	0.005
CP	196401	0.255	0.344	0.877	200301	1.218	0.095

Appendix J Additional In-Sample Results

J.1 Additional Results for LN

Table J.1
The in-sample coefficients and statistics of the additional predictors x_{2t} in the predictive regressions corresponding to LN estimated estimated over the 1985-2016 sample. In each regression, the dependent variable is $\overline{rx}^{(5)}_{t+12}$, x_{1t} consists of a constant and the first three yield curve PCs, and x_{2t} contains the first eight PCs extracted from the extended macroeconomic data set. Under *Wald* the table reports the statistics corresponding to the test of joint significance of the eight macro PCs. The conventional statistics and p-values are computed using Newey-West standard errors with 18 lags. The simple bootstrap procedure is employed. The conventional size and power are estimated using Equation (26) and (29) respectively. The bootstrap 5% critical values (c.v.'s) and p-values are computed using Equation (25) and (24) respectively. The size and power of the bootstrap test are approximated using Equations (27) and (30).

		Stat.	5% c.v.	p-value (in %)		Size (in %)		Power (in %)	
	Coef.	Conv.	Boot.	Conv.	Boot.	Conv.	Boot.	Conv.	Boot.
1985-2016									
f_1	0.955	2.528	3.047	0.012	0.094	0.182	0.056	0.842	0.641
f_2	0.439	2.092	3.078	0.037	0.163	0.187	0.054	0.654	0.407
f_3	-0.389	-1.035	3.259	0.301	0.491	0.203	0.056	0.392	0.161
f_4	0.281	1.267	2.901	0.206	0.356	0.164	0.056	0.381	0.192
f_5	-0.083	-0.320	2.899	0.749	0.813	0.165	0.053	0.195	0.066
f_6	-0.164	-1.341	2.293	0.181	0.252	0.092	0.046	0.309	0.231
f_7	-0.221	-0.742	2.810	0.458	0.586	0.153	0.051	0.308	0.141
f_8	-0.303	-0.955	2.662	0.340	0.467	0.138	0.049	0.469	0.272
Wald		23.071	41.642	0.003	0.259	0.500	0.059	0.947	0.549

J.2 Additional Results for Factor Extraction Methods

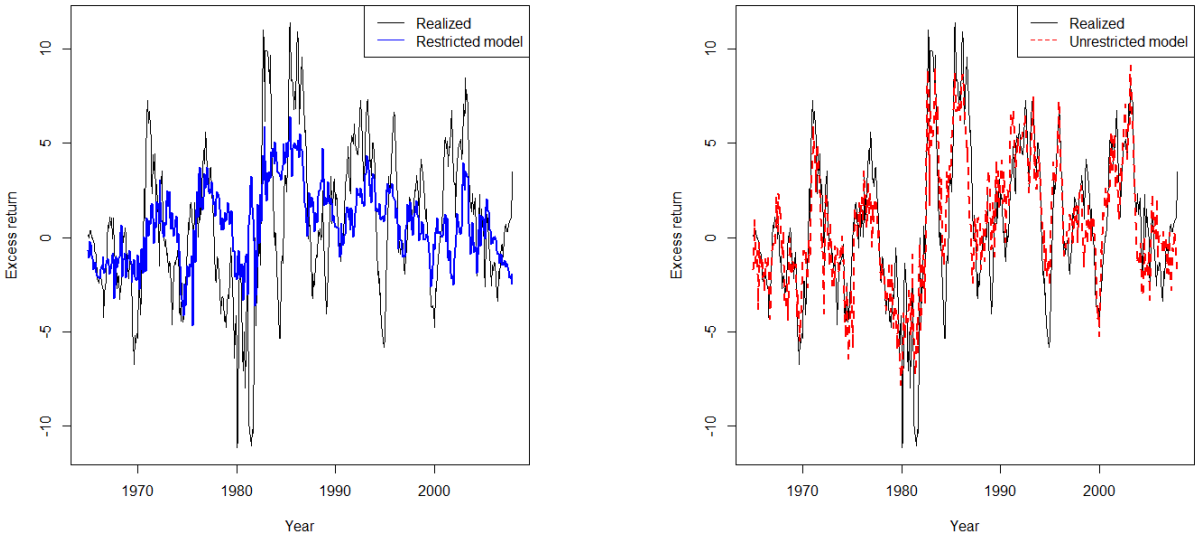
Table J.2
The in-sample coefficients and statistics of the additional predictors x_{2t} in the predictive regressions corresponding to all factor extraction applications estimated over the 1985-2016 sample. In each regression, the dependent variable is $\overline{rx}^{(5)}_{t+12}$, x_{1t} consists of a constant and the first three yield curve PCs, and x_{2t} is the single predictor factor extracted from the extended large macroeconomic data set using the factor extraction methods. The conventional statistics and p-values are computed using Newey-West standard errors with 18 lags. The simple bootstrap procedure is used. The conventional size and power are estimated using Equation (26) and (29) respectively. The bootstrap 5% critical values (c.v.'s) and p-values are computed using Equation (25) and (24) respectively. The size and power of the bootstrap test are approximated using Equations (27) and (30).

			Test Stat.	5% c.v.	p-value		Size	
		Coefficient	Conv.	Bootstrap	Conv.	Bootstrap	Conv.	Bootstrap
No lags								
1964-2007	PCA	0.675	2.984	2.516	0.003	0.023	0.118	0.055
	sPCA	0.705	3.130	2.700	0.002	0.026	0.142	0.049
	FLasso	0.635	2.746	2.418	0.006	0.028	0.108	0.052
	FALasso	0.846	5.806	2.627	0.000	0.000	0.133	0.053
	FCLasso	0.912	4.973	2.301	0.000	0.000	0.098	0.041
	FCALasso	0.782	3.915	2.625	0.000	0.004	0.136	0.053
	FGCALasso1	0.864	4.716	2.569	0.000	0.001	0.124	0.054
	FGCALasso2	0.822	4.194	2.483	0.000	0.002	0.111	0.058
	Lasso	4.336	8.130	2.476	0.000	0.000	0.118	0.050
	ALasso	4.205	9.879	2.451	0.000	0.000	0.111	0.046
	CLasso	5.938	6.617	2.320	0.000	0.000	0.095	0.051
	CALasso	4.402	9.087	2.452	0.000	0.000	0.108	0.052
	GCALasso1	4.072	10.679	2.444	0.000	0.000	0.108	0.049
	GCALasso2	4.043	9.789	2.463	0.000	0.000	0.121	0.045
1985-2022	PCA	1.171	2.818	2.652	0.005	0.040	0.132	0.052
	sPCA	0.909	3.462	2.344	0.001	0.005	0.101	0.057
	FLasso	0.790	3.407	2.362	0.001	0.003	0.096	0.051
	FALasso	0.794	3.414	2.295	0.001	0.005	0.092	0.046
	FCLasso	1.068	3.200	2.257	0.001	0.007	0.086	0.046
	FCALasso	1.056	3.633	2.320	0.000	0.004	0.094	0.050
	FGCALasso1	0.480	0.911	2.785	0.363	0.494	0.152	0.051
	FGCALasso2	0.500	1.157	2.719	0.248	0.376	0.144	0.057
	Lasso	4.212	5.137	2.394	0.000	0.000	0.103	0.050
	ALasso	4.115	3.240	2.414	0.001	0.013	0.106	0.046
	CLasso	8.648	3.735	2.900	0.000	0.017	0.176	0.055
	CALasso	7.879	4.021	2.900	0.000	0.010	0.163	0.057
	GCALasso1	3.121	8.417	2.795	0.000	0.000	0.151	0.055
	GCALasso2	3.114	7.781	2.867	0.000	0.000	0.163	0.056
6 lags								
1964-2007	PCA	0.723	2.981	2.953	0.003	0.048	0.187	0.056
	sPCA	0.748	3.860	2.986	0.000	0.015	0.180	0.048
	FLasso	0.833	5.590	2.497	0.000	0.000	0.126	0.044
	FALasso	0.714	3.611	2.423	0.000	0.004	0.108	0.053
	FCLasso	0.926	5.748	2.454	0.000	0.000	0.113	0.050
	FCALasso	1.036	6.421	2.492	0.000	0.000	0.121	0.048
	FGCALasso1	0.746	3.259	2.505	0.001	0.012	0.116	0.054
	FGCALasso2	0.447	1.406	2.583	0.160	0.284	0.136	0.046
	Lasso	4.900	5.109	2.713	0.000	0.001	0.151	0.058
	ALasso	4.554	11.053	2.598	0.000	0.000	0.129	0.047
	CLasso	6.631	6.927	2.578	0.000	0.000	0.128	0.052
	CALasso	5.013	8.106	2.601	0.000	0.000	0.130	0.054
	GCALasso1	4.274	13.360	2.597	0.000	0.000	0.124	0.056
	GCALasso2	4.248	12.653	2.511	0.000	0.000	0.128	0.049
1985-2022	PCA	1.213	3.373	2.988	0.001	0.029	0.179	0.059
	sPCA	0.891	3.436	3.144	0.001	0.031	0.204	0.062
	FLasso	0.630	3.746	3.291	0.000	0.033	0.210	0.048
	FALasso	0.853	3.631	2.550	0.000	0.007	0.126	0.048
	FCLasso	1.300	2.984	2.666	0.003	0.028	0.136	0.049
	FCALasso	1.063	4.199	2.416	0.000	0.001	0.109	0.052
	FGCALasso1	0.827	3.008	2.545	0.003	0.022	0.123	0.056
	FGCALasso2	0.680	2.826	2.677	0.005	0.038	0.132	0.054
	Lasso	4.817	8.346	2.727	0.000	0.000	0.144	0.047
	ALasso	3.980	8.033	2.662	0.000	0.000	0.140	0.046
	CLasso	14.419	4.058	2.948	0.000	0.011	0.172	0.053
	CALasso	14.184	4.541	2.719	0.000	0.002	0.147	0.049
	GCALasso1	3.525	9.202	3.031	0.000	0.000	0.184	0.056
	GCALasso2	3.069	15.338	3.115	0.000	0.000	0.202	0.060

Table J.3
In-sample adjusted R^2 for the restricted regression model with only x_{1t} (R_1^2), the adjusted R^2 for the unrestricted regression model including both x_{1t} and x_{2t} (R_2^2), and the difference in adjusted R^2 ($R_2^2 - R_1^2$) corresponding to the relevant factor extraction applications estimated over the samples 1964-2007 and 1985-2022. In each regression model, the dependent variable is $\overline{rx}_{t+12}^{(5)}$, x_{1t} consists of a constant and the first three yield curve PCs, and x_{2t} is the single predictor factor extracted from the extended macroeconomic data set with lagged values. No winsorization is applied when extracting the additional predictors. The left half of the table provides the results for the earlier sample periods and the right half of the table provides the results for the later sample periods. For each application, the first row reports the adjusted R^2 statistic in the corresponding actual data set; the second and third rows report respectively the mean and 95%-quantiles of the statistics in the 5,000 bootstrap replications under H_0.

		R_1^2	R_2^2	$R_1^2 - R_2^2$	R_1^2	R_2^2	$R_1^2 - R_2^2$
FEM, 6 lags		Earlier sample, 1964-2007			Later sample, 1985-2022		
PCA	Data	0.25	0.33	0.07	0.16	0.24	0.08
	Bootstrap	0.20	0.22	0.02	0.32	0.33	0.01
		(0.05, 0.39)	(0.07, 0.41)	(-0.00, 0.10)	(0.12, 0.53)	(0.13, 0.54)	(-0.00, 0.07)
sPCA	Data	0.25	0.34	0.08	0.16	0.23	0.07
	Bootstrap	0.20	0.22	0.02	0.32	0.34	0.02
		(0.05, 0.39)	(0.07, 0.41)	(-0.00, 0.11)	(0.12, 0.52)	(0.13, 0.54)	(-0.00, 0.10)
FLasso	Data	0.25	0.40	0.14	0.16	0.18	0.03
	Bootstrap	0.20	0.22	0.01	0.32	0.33	0.01
		(0.05, 0.39)	(0.06, 0.40)	(-0.00, 0.06)	(0.12, 0.53)	(0.13, 0.53)	(-0.00, 0.06)
GCALasso1	Data	0.25	0.76	0.51	0.16	0.52	0.37
	Bootstrap	0.20	0.22	0.01	0.32	0.34	0.02
		(0.06, 0.39)	(0.06, 0.40)	(-0.00, 0.06)	(0.12, 0.52)	(0.14, 0.54)	(-0.00, 0.10)
GCALasso2	Data	0.25	0.72	0.47	0.16	0.74	0.58
	Bootstrap	0.20	0.22	0.01	0.32	0.34	0.02
		(0.05, 0.39)	(0.06, 0.40)	(-0.00, 0.06)	(0.12, 0.53)	(0.14, 0.54)	(-0.00, 0.10)

J.3 Additional In-Sample Plot



(a) Restricted model with only the three yield PCs

(b) Model with the three PCs and GCALasso2 factor

Figure J.1
In-sample forecasts for $\overline{rx}_{t+12}^{(5)}$ resulting from the restricted forecast model with only the three yield curve PCs as predictors (on the left) and the unrestricted model with the three yield curve PCs and the GCALasso2 factor as predictors (on the right) along with the actual excess bond returns over the 1985-2022 sample. The GCALasso2 factor is extracted from the large macroeconomic data set that also includes the six lagged values of the macroeconomic variables. The adjusted R^2 for the restricted regression model is equal to 0.25 and the adjusted R^2 for the unrestricted regression model is equal to 0.72.

J.4 Economic Interpretation

J.4.1 GCALasso2 Factor and Individual Macroeconomic Variables, 1964-2007

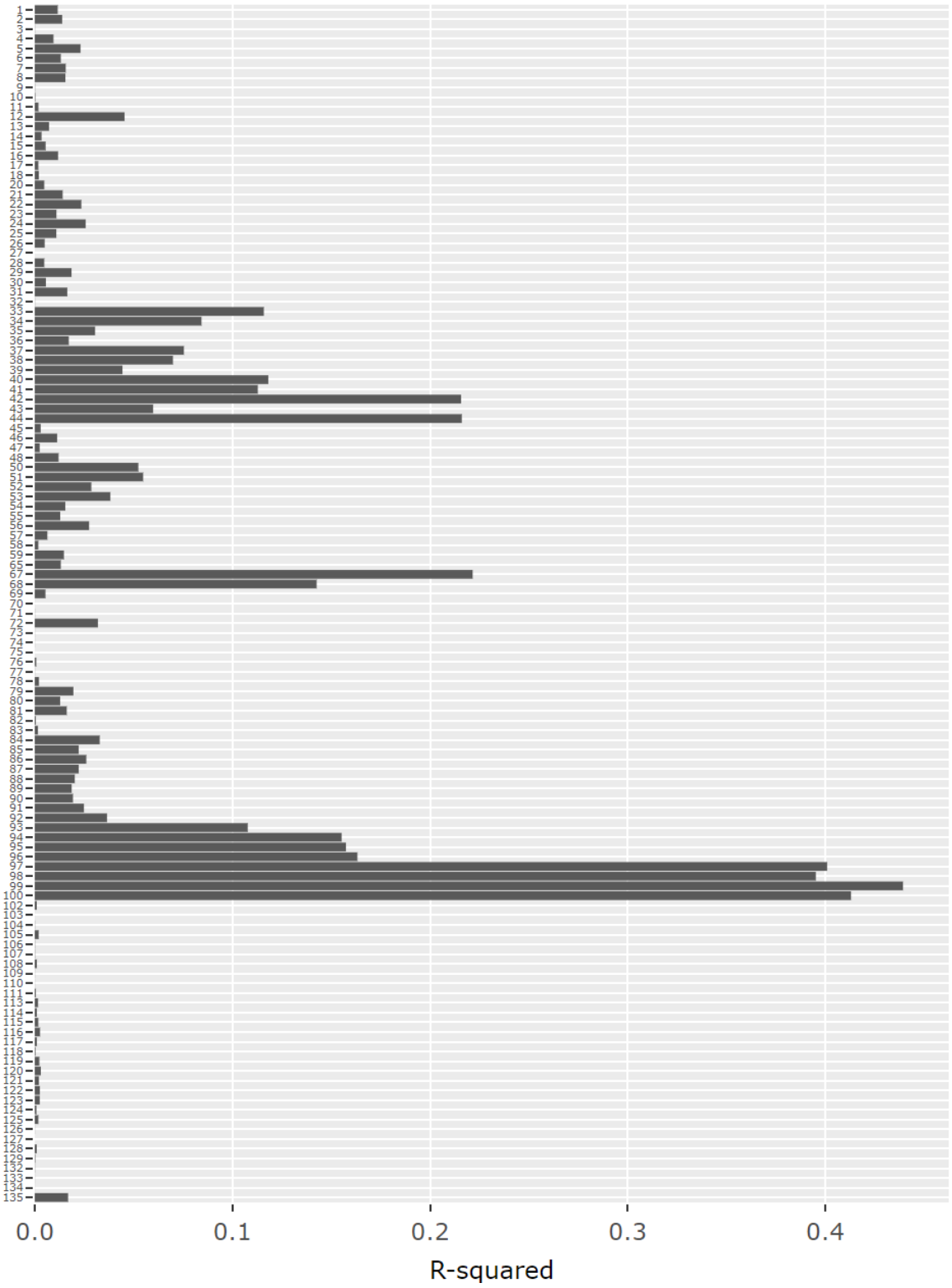


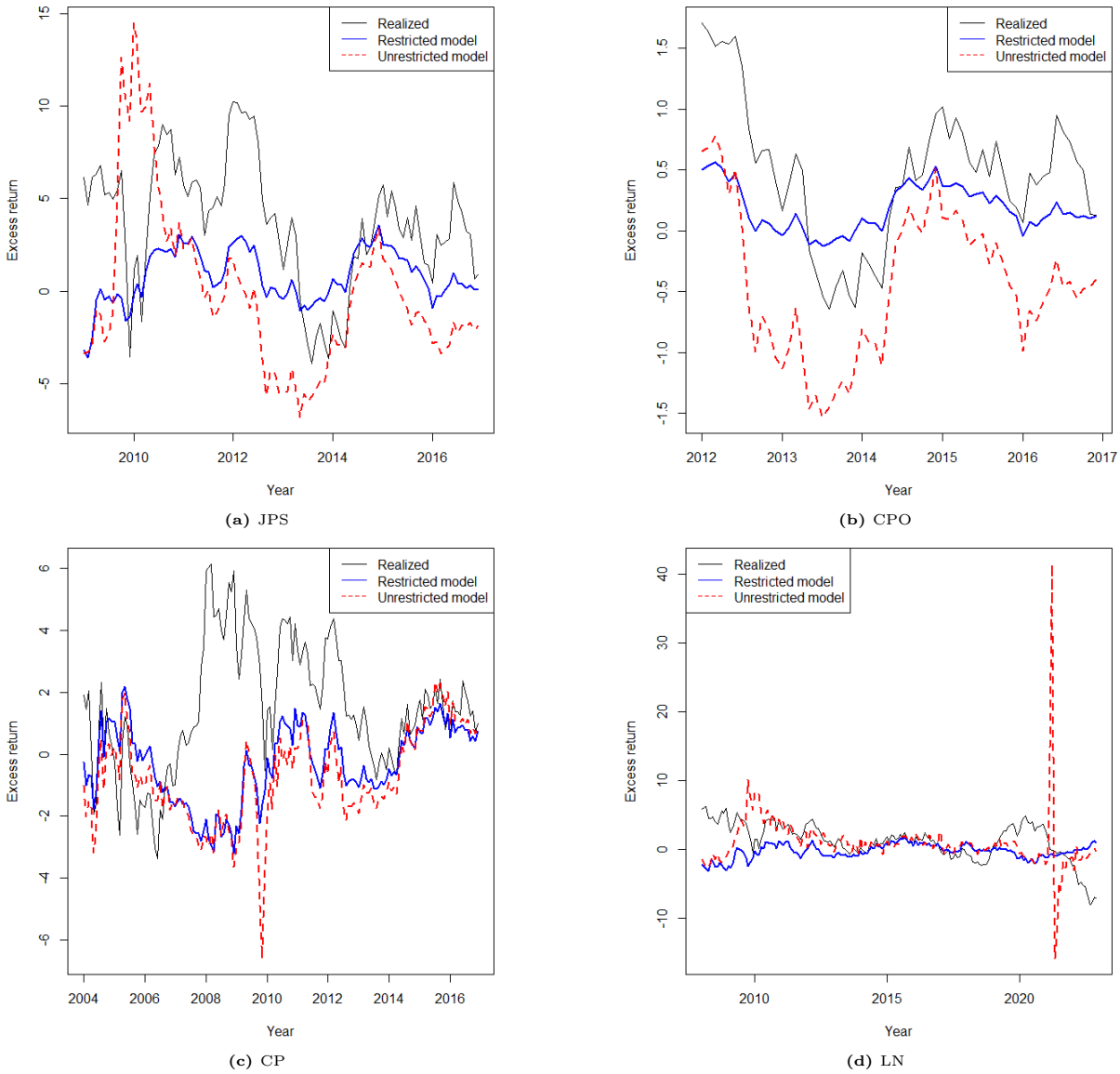
Figure J.2
Marginal R^2 statistics for the regressions of the in-sample GCALasso2 predictor factor on each variable in the large macroeconomic data set. The macroeconomic variables are given on the y-axis and are represented by their ID's. The description of each variable can be found in Table B.1 in Appendix B. The GCALasso2 factor is constructed for the period from January 1964 to December 2007 using the large macroeconomic data set with lagged values.

J.4.2 GCALasso2 Factor and Individual Macroeconomic Variables, 1985-2022

Figure J.3
Marginal R^2 statistics for the regressions of the in-sample GCALasso2 predictor factor on each variable in the large macroeconomic data set. The macroeconomic variables are given on the y-axis and are represented by their ID's. The description of each variable can be found in Table B.1 in Appendix B. The GCALasso2 factor is constructed for the period from January 1985 to December 2022 using the large macroeconomic data set with lagged values.

Appendix K Additional Out-of-Sample Results

K.1 Out-of-Sample Plots for JPS, CPO, CP, LN



(a) JPS **(b)** CPO **(c)** CP **(d)** LN

Figure K.1
Out-of-sample forecasts for average excess bond returns resulting from the restricted forecast model with only the three yield curve PCs as predictors and forecasts of the unrestricted model with the three yield curve PCs and additional predictors as proposed in JPS (a), CPO (b), CP (c), and LN (d). The data sets provided by Bauer and Hamilton (2018) are used to estimate the models for JPS, CPO and CP. The extended data sets are used to estimate the models for LN. The target variable also differs across the applications. It is equal to $\overline{rx}^{(10)}_{t+12}$ for JPS, $\overline{wrx}^{(15)}_{t+12}$ for CPO, and $\overline{rx}^{(5)}_{t+12}$ for CP and LN. Expanding window estimation is used to construct the forecasts. The initial training window is the original sample used in each study. The out-of-sample period starts in the first month after the end of the initial training window and ends in December 2016 for JPS, CPO and CP. For LN, the out-of-sample period ends in December 2022.

K.2 Additional Results for the Factor Extraction Methods

Table K.1
Out-of-sample predictive power for $\overline{rx}^{(5)}_{t+12}$ of a restricted model with the three yield curve PCs and an unrestricted model with additional predictors corresponding to all the relevant factor extraction methods in different scenarios. The results are shown for the scenarios in which no winsorization, input winsorization or output winsorization is applied. Furthermore, in case winsorization is applied, the results are also shown for the scenarios in which the factors are extracted from the large macroeconomic data set with lagged values. The additional predictors x_{2t} for the PCA, sPCA, FLasso, FALasso, FCLasso, FCALasso, FGCALasso1 and FGCALasso2 methods are either the six factors or the single predictor factor resulting from these methods. Since the other methods only produce a single predictor factor, the additional predictor x_{2t} for these methods is the factor resulting from each of these two methods. The in-sample period starts in January 1964 and ends in December 2007. The out-of-sample period starts one month later than the end of the in-sample period and ends in either December 2016 or December 2022. To generate the out-of-sample forecasts, expanding window estimation is used. Under *MSE ratio* and *p-value* the table reports respectively the mean-squared errors for the unrestricted model relative to the mean-squared errors for the restricted model and the p-values of the Diebold-Mariano test for equal prediction accuracy of the two models.

	Individual Factors				Single Joint Factor			
	End: 2016		End: 2022		End: 2016		End: 2022	
	MSE ratio	p-value	MSE ratio	p-value	MSE ratio	p-value	MSE ratio	p-value
No winsorization and no lags								
PCA	0.753	0.171	1.098	0.786	0.776	0.231	1.431	0.526
sPCA	0.821	0.389	1.444	0.486	0.809	0.664	1.047	0.891
FLasso	0.861	0.620	1.250	0.562	0.894	0.742	1.329	0.489
FALasso	0.710	0.453	1.056	0.868	0.783	0.555	1.118	0.719
FCLasso	0.740	0.408	1.295	0.587	0.846	0.712	1.461	0.476
FCALasso	0.768	0.462	1.675	0.421	0.889	0.717	1.474	0.441
FGCALasso1	0.865	0.567	1.572	0.432	0.824	0.327	1.467	0.468
FGCALasso2	0.788	0.547	1.652	0.459	0.915	0.749	1.519	0.421
Lasso					1.224	0.661	1.647	0.265
ALasso					1.151	0.767	1.506	0.261
CLasso					0.798	0.530	1.161	0.675
CALasso					1.164	0.579	1.596	0.299
GCALasso1					1.333	0.167	1.909	0.166
GCALasso2					1.442	0.225	2.021	0.194
Input winsorization and no lags								
PCA	0.584	0.054	0.656	0.024	0.617	0.062	0.647	0.018
sPCA	0.701	0.083	0.733	0.058	0.523	0.113	0.618	0.061
FLasso	0.466	0.109	0.600	0.061	0.579	0.124	0.690	0.073
FALasso	0.491	0.138	0.627	0.085	0.630	0.151	0.724	0.079
FCLasso	0.518	0.145	0.675	0.143	0.521	0.108	0.645	0.076
FCALasso	0.547	0.112	0.685	0.103	0.596	0.129	0.695	0.081
FGCALasso1	0.666	0.133	0.736	0.100	0.644	0.111	0.734	0.090
FGCALasso2	0.641	0.078	0.744	0.107	0.629	0.106	0.689	0.043
Lasso					0.956	0.786	0.919	0.592
ALasso					0.859	0.533	0.900	0.564
CLasso					0.520	0.119	0.652	0.098
CALasso					0.854	0.375	0.879	0.422
GCALasso1					1.256	0.188	1.108	0.521
GCALasso2					1.364	0.169	1.168	0.386
Output winsorization and no lags								
PCA	0.560	0.073	0.643	0.030	0.572	0.051	0.673	0.037
sPCA	0.577	0.069	0.671	0.041	0.438	0.130	0.609	0.086
FLasso	0.527	0.059	0.652	0.037	0.527	0.039	0.702	0.058
FALasso	0.495	0.183	0.641	0.118	0.676	0.390	0.840	0.465
FCLasso	0.519	0.158	0.681	0.153	0.475	0.128	0.654	0.109
FCALasso	0.540	0.190	0.747	0.283	0.591	0.135	0.758	0.178
FGCALasso1	0.595	0.134	0.716	0.112	0.714	0.173	0.809	0.167
FGCALasso2	0.517	0.128	0.712	0.200	0.619	0.085	0.759	0.103
Lasso					0.786	0.345	0.880	0.455
ALasso					0.834	0.650	0.974	0.909
CLasso					0.515	0.121	0.697	0.139
CALasso					0.812	0.316	0.842	0.294
GCALasso1					1.152	0.521	1.173	0.303
GCALasso2					1.145	0.581	1.103	0.546

Continued on next page

Continued from previous page

	Individual Factors				Single Joint Factor			
	End: 2016		End: 2022		End: 2016		End: 2022	
	MSE ratio	p-value	MSE ratio	p-value	MSE ratio	p-value	MSE ratio	p-value
Input winsorization and 6 lags								
PCA	0.555	0.089	0.609	0.042	0.572	0.086	0.609	0.024
sPCA	0.753	0.132	0.748	0.039	0.566	0.192	0.620	0.074
FLasso	0.544	0.240	0.653	0.134	0.510	0.286	0.657	0.192
FALasso	0.638	0.176	0.722	0.107	0.625	0.105	0.746	0.080
FCLasso	0.742	0.583	0.783	0.428	0.615	0.303	0.677	0.144
FCALasso	0.690	0.271	0.736	0.149	0.521	0.212	0.625	0.101
FGCALasso1	0.682	0.367	0.767	0.291	0.559	0.267	0.643	0.131
FGCALasso2	0.645	0.332	0.732	0.233	0.459	0.211	0.584	0.104
Lasso					0.558	0.392	0.667	0.259
ALasso					0.967	0.951	0.920	0.806
CLasso					0.563	0.350	0.658	0.218
CALasso					1.150	0.738	1.012	0.967
GCALasso1					1.110	0.754	1.014	0.950
GCALasso2					0.927	0.858	0.876	0.614
Output winsorization and 6 lags								
PCA	0.541	0.087	0.610	0.029	0.525	0.078	0.617	0.035
sPCA	0.597	0.118	0.711	0.080	0.453	0.253	0.601	0.160
FLasso	0.433	0.142	0.645	0.127	0.297	0.142	0.581	0.150
FALasso	0.581	0.093	0.745	0.106	0.573	0.121	0.752	0.151
FCLasso	0.538	0.262	0.661	0.168	0.496	0.217	0.655	0.151
FCALasso	0.523	0.195	0.661	0.133	0.488	0.158	0.643	0.102
FGCALasso1	0.738	0.221	0.775	0.123	0.664	0.102	0.717	0.031
FGCALasso2	0.825	0.455	0.923	0.604	0.682	0.144	0.778	0.089
Lasso					0.318	0.189	0.566	0.161
ALasso					0.588	0.304	0.772	0.347
CLasso					0.391	0.136	0.569	0.083
CALasso					0.404	0.155	0.593	0.106
GCALasso1					1.004	0.986	0.922	0.619
GCALasso2					1.021	0.957	0.973	0.906

K.3 Economic Interpretation

K.3.1 PCA Factor and Individual Macroeconomic Variables, 2007-2021

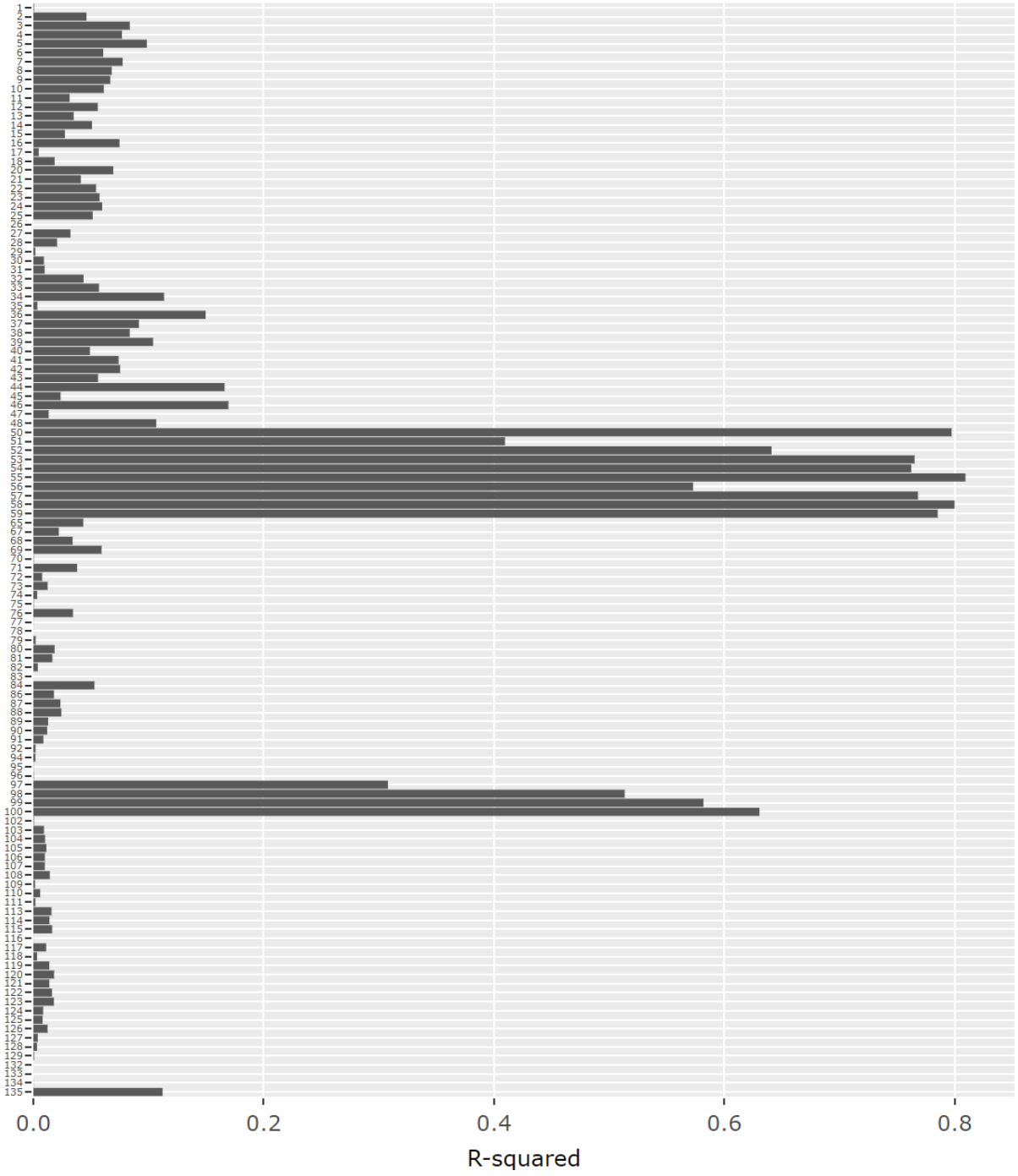


Figure K.2
Marginal R^2 statistics for the regressions of the out-of-sample PCA single predictor factor on each variable in the large macroeconomic data set. The macroeconomic variables are given on the y-axis and are represented by their ID's. The description of each variable can be found in Table B.1 in Appendix B. The PCA factor is calculated for the period from January 2007 to December 2021 using expanding window estimation. The training windows start in January 1964. Output winsorization is applied recursively.

K.3.2 GCALasso2 Factor and Individual Macroeconomic Variables, 2007-2021

Figure K.3
Marginal R^2 statistics for the regressions of the out-of-sample GCALasso2 predictor factor on each variable in the large macroeconomic data set. The macroeconomic variables are given on the y-axis and are represented by their ID's. The description of each variable can be found in Table B.1 in Appendix B. The GCALasso2 factor is extracted from the large data set containing the macroeconomic variables and their six lagged values and is calculated for the period 2007-2021 using expanding window estimation. The training windows start in January 1964. Output winsorization is applied recursively

K.4 Robustness Checks

In this section, I investigate the sensitivity of the out-of-sample results with respect to the number of predictor factors and the winsorization thresholds. In the sensitivity analysis, I focus on the most relevant factor extraction methods, namely the PCA, sPCA, FLasso, GCALasso1 and GCALasso2 methods. Since adding lagged values to the large macroeconomic data set generally leads to better results for the GCALasso1 and GCALasso2 methods, the predictor factors corresponding to these methods are extracted from the data set with lagged values. The predictor factors corresponding to the other three methods are extracted from the large macroeconomic data set without lagged values. Furthermore, I ignore the case in which no winsorization is applied, as it has been shown that winsorization helps to substantially improve the out-of-sample results.

Following D. Huang et al. (2023), six factors are extracted with the methods involving PCA. However, one may be concerned that the results are dependent on this particular choice. To investigate this, I also generate the results for the cases in which four, five, seven or eight factors are extracted instead of six. The corresponding results for the PCA, sPCA and FLasso methods are given in Table K.2. In general, the performance of the methods is quite similar if alternative numbers of factors are extracted. However, the evidence that the unrestricted forecasting model outperforms the restricted one is slightly weaker when extracting seven or eight factors. For the sPCA method, performance improves when the number of factors decreases. In the case of four factors, the reduction in MSE is considerably larger for the sPCA method than for the PCA method in all scenarios, but the reduction is not significant at 10% level in all scenarios for the sPCA method while it is for the PCA method. The PCA method thus leads to the most consistent results in terms of significance.

Following Bottmer et al. (2022), I applied 90% winsorization in the empirical analysis. To test the robustness of the results with respect to this decision, I also consider 80%, 95% and 99% winsorization. The results for the PCA, sPCA, FLasso, GCALasso1 and GCALasso2 methods are given in Table K.3. The performance of the methods slightly improves when the winsorization thresholds are smaller, but overall the performance is comparable across the different winsorization intervals. With two exceptions in the case of 99% winsorization, the PCA method leads to significant improvements at 10% level in all scenarios. Adding either the GCALasso1 or GCALasso2 factor to the restricted forecasting model leads to worse forecasts in all scenarios. Overall, it can be concluded that using the alternative winsorization thresholds leads to similar conclusions.

Table K.2
Out-of-sample predictive power for $\overline{rx}_{t+12}^{(5)}$ of a restricted model with the three yield curve PCs and an unrestricted model with both the three yield PCs and either multiple individual factors or the single predictor factor resulting from the relevant factor extraction methods for different scenarios. The results are shown for the scenarios in which four, five, six, seven or eight factors are extracted and either input winsorization or output winsorization is applied. The in-sample period starts in January 1964 and ends in December 2007. The out-of-sample period starts one month later than the end of the in-sample period and ends in either December 2016 or December 2022. To generate the out-of-sample forecasts, expanding window estimation is used. Under *MSE ratio* and *p-value* the table reports respectively the mean-squared errors for the unrestricted model relative to the mean-squared errors for the restricted model and the p-values of the Diebold-Mariano test for equal prediction accuracy of the two models.

	Individual Factors				Single Joint Factor			
	End: 2016		End: 2022		End: 2016		End: 2022	
	MSE ratio	p-value	MSE ratio	p-value	MSE ratio	p-value	MSE ratio	p-value
4 factors								
Input winsorization								
PCA	0.586	0.073	0.654	0.029	0.627	0.055	0.651	0.015
sPCA	0.502	0.087	0.596	0.035	0.535	0.100	0.589	0.031
FLasso	0.548	0.109	0.647	0.051	0.713	0.160	0.765	0.074
Output winsorization								
PCA	0.546	0.067	0.621	0.022	0.559	0.046	0.645	0.020
sPCA	0.422	0.108	0.553	0.048	0.443	0.101	0.577	0.052
FLasso	0.568	0.066	0.692	0.043	0.640	0.114	0.802	0.200
5 factors								
Input winsorization								
PCA	0.580	0.057	0.656	0.025	0.625	0.064	0.654	0.018
sPCA	0.581	0.067	0.639	0.026	0.515	0.120	0.578	0.039
FLasso	0.508	0.116	0.654	0.075	0.611	0.116	0.742	0.097
Output winsorization								
PCA	0.553	0.069	0.629	0.024	0.570	0.046	0.658	0.029
sPCA	0.576	0.075	0.681	0.043	0.415	0.110	0.586	0.069
FLasso	0.561	0.077	0.680	0.048	0.651	0.121	0.791	0.173
6 factors								
Input winsorization								
PCA	0.584	0.054	0.656	0.024	0.617	0.062	0.647	0.018
sPCA	0.701	0.083	0.733	0.058	0.523	0.113	0.618	0.061
FLasso	0.466	0.109	0.600	0.061	0.579	0.124	0.690	0.073
Input winsorization								
PCA	0.560	0.073	0.643	0.030	0.572	0.051	0.673	0.037
sPCA	0.577	0.069	0.671	0.041	0.438	0.130	0.609	0.086
FLasso	0.527	0.059	0.652	0.037	0.527	0.039	0.702	0.058
7 factors								
Input winsorization								
PCA	0.636	0.104	0.688	0.050	0.651	0.154	0.669	0.051
sPCA	0.720	0.116	0.752	0.072	0.538	0.131	0.632	0.074
FLasso	0.532	0.116	0.660	0.089	0.648	0.170	0.758	0.150
Output winsorization								
PCA	0.557	0.085	0.630	0.033	0.564	0.088	0.631	0.037
sPCA	0.562	0.072	0.664	0.038	0.416	0.142	0.585	0.086
FLasso	0.532	0.135	0.687	0.115	0.549	0.126	0.735	0.163
8 factors								
Input winsorization								
PCA	0.612	0.106	0.675	0.052	0.640	0.152	0.661	0.050
sPCA	0.763	0.162	0.772	0.058	0.536	0.136	0.624	0.061
FLasso	0.555	0.176	0.692	0.156	0.647	0.243	0.775	0.241
Output winsorization								
PCA	0.555	0.085	0.624	0.036	0.567	0.093	0.655	0.051
sPCA	0.540	0.081	0.645	0.039	0.377	0.151	0.563	0.097
FLasso	0.517	0.134	0.693	0.135	0.504	0.139	0.753	0.259

Table K.3
Out-of-sample predictive power for $\overline{rx}^{(5)}_{t+12}$ of a restricted model with the three yield curve PCs and an unrestricted model with additional predictors corresponding to the relevant factor extraction methods in different scenarios. The results are shown for the scenarios in which 80%, 90%, 95% or 99% winsorization is applied. The additional predictors x_{2t} for the LN application are either the eight macro PCs or a single predictor factor which is the fitted value of the regression of $\overline{rx}^{(5)}_{t+12}$ on the eight macro PCs. The additional predictors x_{2t} for the PCA, sPCA and FLasso methods are either the six factors or the single predictor factor resulting from these methods. Since the GCALasso1 and GCALasso2 methods only produce a single predictor factor, the additional predictor x_{2t} for these methods is the factor resulting from each of these two methods. The in-sample period starts in January 1964 and ends in December 2007. The out-of-sample period starts one month later than the end of the in-sample period and ends in either December 2016 or December 2022. To generate the out-of-sample forecasts, expanding window estimation is used. Under *MSE ratio* and *p-value* the table reports respectively the mean-squared errors for the unrestricted model relative to the mean-squared errors for the restricted model and the *p*-values of the Diebold-Mariano test for equal prediction accuracy of the two models. The *p*-values that are lower than 5% are highlighted in bold.

	Individual Factors				Single Joint Factor			
	End: 2016		End: 2022		End: 2016		End: 2022	
	MSE ratio	*p*-value	MSE ratio	*p*-value	MSE ratio	*p*-value	MSE ratio	*p*-value
80% winsorization								
Input winsorization								
PCA	0.576	0.083	0.655	0.038	0.647	0.088	0.666	0.027
sPCA	0.661	0.099	0.714	0.062	0.517	0.104	0.620	0.065
FLasso	0.549	0.141	0.653	0.083	0.604	0.131	0.715	0.092
GCALasso1					1.256	0.272	1.148	0.332
GCALasso2					1.222	0.320	1.120	0.408
Output winsorization								
PCA	0.547	0.085	0.624	0.032	0.623	0.076	0.701	0.044
sPCA	0.549	0.064	0.641	0.034	0.472	0.172	0.636	0.120
FLasso	0.480	0.072	0.631	0.047	0.493	0.083	0.676	0.082
GCALasso1					1.067	0.681	1.116	0.381
GCALasso2					1.156	0.539	1.156	0.354
90% winsorization								
Input winsorization								
PCA	0.584	0.054	0.656	0.024	0.617	0.062	0.647	0.018
sPCA	0.701	0.083	0.733	0.058	0.523	0.113	0.618	0.061
FLasso	0.466	0.109	0.600	0.061	0.579	0.124	0.690	0.073
GCALasso1					1.256	0.188	1.108	0.521
GCALasso2					1.364	0.169	1.168	0.386
Output winsorization								
PCA	0.560	0.073	0.643	0.030	0.572	0.051	0.673	0.037
sPCA	0.577	0.069	0.671	0.041	0.438	0.130	0.609	0.086
FLasso	0.527	0.059	0.652	0.037	0.527	0.039	0.702	0.058
GCALasso1					1.152	0.521	1.173	0.303
GCALasso2					1.145	0.581	1.103	0.546
95% winsorization								
Input winsorization								
PCA	0.607	0.049	0.677	0.023	0.629	0.057	0.660	0.016
sPCA	0.724	0.078	0.743	0.049	0.547	0.163	0.625	0.073
FLasso	0.626	0.070	0.746	0.068	0.699	0.154	0.782	0.107
GCALasso1					1.138	0.144	1.057	0.686
GCALasso2					1.236	0.212	1.133	0.416
Output winsorization								
PCA	0.595	0.064	0.678	0.030	0.568	0.045	0.683	0.043
sPCA	0.620	0.081	0.706	0.051	0.464	0.134	0.627	0.089
FLasso	0.576	0.093	0.691	0.063	0.521	0.106	0.697	0.106
GCALasso1					1.128	0.501	1.121	0.361
GCALasso2					1.218	0.469	1.191	0.321
99% winsorization								
Input winsorization								
PCA	0.698	0.086	0.753	0.057	0.717	0.120	0.731	0.040
sPCA	0.815	0.300	0.813	0.154	0.704	0.434	0.720	0.221
FLasso	0.682	0.423	0.730	0.247	0.818	0.718	0.850	0.593
GCALasso1					1.219	0.300	1.073	0.676
GCALasso2					1.165	0.601	1.038	0.856
Output winsorization								
PCA	0.693	0.072	0.749	0.041	0.696	0.089	0.786	0.138
sPCA	0.769	0.245	0.808	0.149	0.651	0.354	0.750	0.260
FLasso	0.779	0.469	0.825	0.350	0.715	0.362	0.821	0.336
GCALasso1					1.224	0.253	1.181	0.182
GCALasso2					1.367	0.267	1.296	0.163

Appendix L Programmed Code

The ZIP file 'Data&Codes.zip' includes the following programs.

- cp.r - Generates the in-sample results for the CP application.
- cpo.r - Generates the in-sample results for the CPO application.
- cpo_1m.r - Generates additional in-sample results for the CPO application in the cases without overlapping returns (using monthly returns).
- cpo_figure.r - Generates the plot with the 10-yield and inflation trend over time for the CPO application.
- economic_interpretation.r - Generates bar plots and statistics used to interpret the in-and-out-of-sample GCALasso2 factor and the out-of-sample PCA factor economically.
- fem_is.r - Generates the basic in-sample results for the factor extraction methods (FEM).
- fem_lags_is.r - Generates the in-sample results for the factor extraction methods (FEM) applied to the large macroeconomic data set with lagged values.
- fem_oos.r - Generates the out-of-sample results for the LN application and the other factor extraction methods (FEM).
- figure_extreme_values.r - Generates the plots that illustrate the extreme values in the large macroeconomic data set.
- figure_fit_GCALasso2_is.r - Generates the plots that show the in-sample fit the restricted model with only the three yield PCs and the unrestricted model with both the three yield PCs and the GCALasso2 factor.
- jps.r - Generates in-sample results for the JPS application.
- ln.r - Generates in-sample results for the LN application.
- rev_oos.r - Generates the out-of-sample results for the revisited studies.
- robustness_numFactors_fem_oos.r - Generates the out-of-sample results for the scenario in which the user-specified number of factors are extracted using the relevant factor extraction methods (FEM).
- robustness_winsorizationThresholds_fem_oos.r - Generates the out-of-sample results for the relevant factor extraction methods (FEM) in the scenario in which winsorization is applied with the user-specified winsorization thresholds.

- sim.r - Generates the simulation results for the case without overlapping observations.
- sim_overlapping_returns.r - Generates the simulation results for the case without overlapping observations.
- sim_size_T.r - Generates the simulation results that show the size distortions across different sample sizes.
- R\robust_fns.r - Contains additional functions that are used in the scripts listed above. These functions are loaded when the scripts are run.
- R\var_fns.r - Contains functions written by Bauer and Hamilton (2018) to estimate the VAR models in the bootstrap procedure.