An Affine Term Structure Model With a Large Number of Factors and Machine Learning

Ga-Young Jang
Department of Finance, Hanyang University Business School, 222 Wangsimni-ro, Seongdong-gu, Seoul, Korea 04763; 1 858 209 3515; kathy.g.jang@gmail.com

Hyoung-Goo Kang (C.A.)
Department of Finance, Hanyang University Business School, 222 Wangsimni-ro, Seongdong-gu, Seoul, Korea 04763; +82 2 2220 1177; hyoungkang@hanyang.ac.kr

Daejin Kim
Department of Fintech, Sungkyunkwan University, 25-2, Sungkyunkwan-ro, Jongno-gu, Seoul, Korea 03063; +82 2 760 0847; daejin@skku.edu

Dongjoon Lee
Department of Fintech, Sungkyunkwan University, 25-2, Sungkyunkwan-ro, Jongno-gu, Seoul, Korea 03063; +82 10 2113 8119; chasae94@gmail.com

Abstract

This study investigates a new affine term structure model that uses machine learning methods in addition to traditional principal component and latent factor analysis. Prior asset pricing models have limitations due to small sample sizes, computational complexity, and lack of economic implications, to which the proposed model provides a partial solution. We use a regression method and PCA to construct the model with five yield factors and 124 macroeconomic variables. The empirical results support the model's robust yield fitting and predictability for future excess returns. The model's predictability further improves as the number of factors increases. Using Lasso for data augmentation, we exemplify how the model can use machine learning to expand data infinitely for greater accuracy in predicting bond returns. We also use Lasso when selecting the most relevant macroeconomic variables to be used in the prediction of bond returns. Our findings extend the literature by showing how to apply machine learning without breaching the no-arbitrage rule and provide fixed income managers with a convenient tool to predict bond returns while generating economic interpretations.

Keywords

Affine term structure model; bond pricing; machine learning; macroeconomic factors

JEL classifications

G12 E43 E44

I. Introduction

Early asset pricing models, such as the capital asset pricing model (CAPM) (Sharpe, 1964) and the Intertemporal CAPM (ICAPM) (Merton, 1973), focus on predicting asset returns using observable factors. Although such models provide statistical and systematic means to predict asset returns for academics and practitioners, using observable factors limits the scope of investigation or interpretation of the results. In 1964, the arbitrage pricing theory (APT) (Ross, 1976) introduces the use of latent factors, building the link between returns and risk premia although latent factor analysis also faces shortcomings. For example, while macroeconomic variables can be used as inputs in model construction, the results cannot be linked to specific economic implications (Bai & Wang, 2016).

Over the past decades, despite such shortcomings, the traditional factor models have gained popularity among academics and practitioners in predicting asset returns, especially for stock investment. However, the models can hardly be used for predicting bond returns. It is mainly because bonds are issued in different tranches and maturities, which affect the price movement of different bonds upon issuance. This highlights the importance of investigating a yield curve, which plots the cross-sectional relationship between maturities and yields of bonds in a given time. Therefore, academics have developed affine models to explore this relationship and predict future bond returns in particular. For example, Litterman and Scheinkman (1991) use a principal component analysis (PCA) and find three principal components, such as level, slope, and curvature factors, as main determinants of US bond returns. Since then, affine models have been the main focus of bond pricing studies, some of which characterize affine models using latent factors (Duffie & Kan, 996), investigate an additional return-forecasting factor as the

fourth principal component (Cochrane & Piazzesi, 2008), and use a computationally simple regression method to construct an affine model (Adrian, Crump, & Moench, 2013).

Nevertheless, due to the computational burden and restrictions imposed on affine models, existing studies seldom (i) extend the number of factors to find the most precise specification for the model or (ii) explore innovative methods to extend the model's use. Besides, (iii) a small sample bias also exists. In terms of the number of factors, asset pricing models for equities explore up to 316 factors (Harvey, Liu, and Zhu, 2016) to investigate its return predictability. For bonds, the most up-to-date innovation for affine term structure model literature may be the discovery of the five principal factors through the regression method (Adrian, et al., 2013). Affine models before the ACM model primarily used maximum likelihood methods to estimate principal factors (Chen & Scott, 1993). By using regressions, the ACM model incorporates higher-order factors such as the fourth and fifth principal components without compromising the model's predictive power or increasing computational difficulty. However, the other two problems still remain, that is, the model's inability to provide economic interpretations and a small sample bias.

For innovative methods, some existing studies investigate the use of machine learning in asset pricing model construction even for bonds (Bianchi, Buchner, & Tamoni, 2020). Machine learning methods can also be used to address the small sample issue. However, studies that utilize machine learning for asset pricing are subject to criticism such that they seldom base their investigation on important asset pricing assumptions, i.e., no arbitrage. The use of machine learning in asset pricing is being extensively discussed by academics and practitioners (Israel et al., 2020), but how to address it while adhering to the traditional assumption of no arbitrage, for example, may be another challenge that needs to be solved first.

To address such concerns and fill the gaps in the literature, this study investigates a new asset pricing model that has the following advantages. First, the new model can incorporate an infinite number of factors without compromising the model's predictability or increasing computational difficulty. To do so, we combine machine learning and regression methods in model construction as well as empirical tests. The regression method we use in this study can be considered a variant of the ACM model.

Furthermore, the proposed model allows the incorporation of an extensive pool of macroeconomic variables. This makes simulations easier, enabling the interpretation of the results with specific economic data the users choose to apply. Macroeconomic variables are essential for predicting bond returns. Prior studies not only shed light on the significant effect of the changes in macroeconomic variables on bond returns (Estrella & Hardouvelis, 1991) but also investigate specific types of macroeconomic variables that affect bond returns (Ludvigson & Ng, 2009; Wachter, 2006); however, the investigation of the influence of macroeconomic variables has been limited in scope due to the prior models' limited capacity to accommodate a large number of factors. We show how to alleviate this issue using the Least Absolute Shrinkage and Selection Operator (Lasso) in addition to PCA.

Lastly, the proposed model easily accommodates machine learning. Finance and economic researchers use high-dimensional data sets characterized by N, a large cross-sectional dimension, and T, a large time dimension. By utilizing machine learning, one can expand the data size by increasing N in particular without limitation. In this study, we exemplify the use of Lasso to expand data and confirm the original results obtained through the ordinary least square (OLS) regressions. Autoencoder, a deep learning model, is one of the other options that can be used along with Lasso and PCA to improve the performance of our model.

In terms of methodology, we first show how to construct our data and model. We conduct PCA on the combined data of US treasury data and a pool of macroeconomic factors to generate latent pricing factors for model construction. The model generates forward rates, which are compared to the observed forward rates during the sample period of 1971:8 to 2021:2 to confirm fitted yields. We run regressions with four to 129 factors, including 124 macroeconomic variables, to investigate the model's forecasting power for future bond returns. To check the robustness of the results, we run Lasso regression with three different specifications using in-sample and out-of-sample data. In the process, we describe how to expand data size infinitely using Lasso. The results confirm that the predictive power of the model remains intact or even improves for intermediate maturities in particular.

Throughout the process, we compare the yield fitting and predictability of our model to those of the ACM model with four or five factors. The outcomes provide strong evidence supporting the predictability of our proposed model for predicting future bond returns with minimal root mean square errors (RMSEs) among other measures that we present alongside. Finally, we perform Lasso in addition to PCA to narrow the list of macroeconomic variables to contain only the relevant ones. The final list contains 23 macroeconomic variables out of 124 variables, which would allow us to interpret the results with more specific economic implications.

Our contributions to the literature are summarized as follows. First, we extend the asset pricing literature by introducing an innovative model, yet strictly grounded on the traditional assumption of asset pricing, i.e., no arbitrage. We propose that the model can help mitigate the problems associated with the computational complexity, limited use of factors, and lack of economic meanings of the existing factor models.

Second, we marginally add to the academic discussion regarding the use of machine learning in asset pricing by proposing a model that has flexibility in accommodating machine learning. We show how the use of machine learning can help expand underlying data and incorporate any number of macroeconomic variables to increase the predictability of the model. When predicting bond returns, for example, fixed income managers tend to consider the movement of a set of macroeconomic variables they consider important and investigate their influence on returns. Since using latent factors, one cannot pinpoint exactly what macroeconomic variables contribute to the results; however, one can at least test different sets of variables with much flexibility and generate economic meanings using our model.

The rest of the paper is organized as follows: Section 2 discusses affine models in general and introduces our new model. Section 3 explains the data and empirical methods used in our study. Section 4 reports the empirical results and Section 5 concludes.

2. Affine models

2.1. Literature review

The proposed model can be considered a variant of that invented by Adrian, Crump, and Moench (ACM) (2013) in the sense that both models use a regression method. Using a regression method entails a number of advantages over other methods such as maximum likelihood (ML) methods, which make distributional assumptions and computation more difficult.

Traditional factor analysis well establishes that the level, slope, and curvature factors drive yields. Cochrane and Pizzesi (2009) discover the single return-forecasting factor next to the traditional three factors, followed by Adrian, et al. (2013) who investigate an affine model with

even higher order factors; however, compared to the asset pricing literature on equities, the asset pricing literature on fixed income assets has more room for development.

One of the reasons is related to the complex nature of bond pricing due to the wide array of features that make each bond unique. Specifically, besides the characteristics specific to issuers, bond maturities or other covenants such as collateralizations or call options affect bond pricing to different degrees. More importantly, the treasury yield curve at the time of issuance is a major determining factor in bond pricing while changes in macroeconomic factors significantly determine the yield curve. For example, bond risk premia are significantly driven by shocks to inflation and aggregate consumption (Brandt and Wang, 2003; Wachter, 2006) or even counter-cyclical movements caused by macroeconomic uncertainty (Bansal & Yaron, 2004; Bansal et al., 2005). Ludvigson and Ng (2009), who use more than a hundred macroeconomic indicators to investigate the effect of cyclical fluctuations of them in bond pricing, criticize that existing affine models are constructed with non-cyclical financial factors only, which does not truly reflect reality.

Nevertheless, affine models up to date can hardly be used to link economic interpretations to their outcomes mainly because they depend on latent factors. The use of latent factors for bond pricing simplifies the process of model construction but also makes it difficult to interpret the results from an economic perspective. To overcome such problems, one of the most simple ways would be to make it computationally easier and faster to test different sets of macroeconomic variables so that users of the model can find the best possible set of macroeconomic variables for predicting future bond returns. Our proposed model makes these possible and allows machine learning methods to be employed without foregoing the no-arbitrage rule.

2.2. The proposed model

This study examines how to construct and test an affine model with a large number of macroeconomic variables as inputs in addition to traditional yield factors. As follows, we describe our empirical strategy step by step as follows. A complete description of the model is included in Appendix I.

Step 1: Normalize Liu and Wu's (2021) yield curve data for the mean value of zero and the standard deviation of one before performing PCA to generate five yield factors.

Step 2: Download the monthly macroeconomic data from FRED. Then, use the *tcode* from McCracken and Ng (2015) for data transformation. Following the method described in Appendix II, transform the unbalanced panel of data to a balanced one.

Step 3: Perform another PCA on the obtained 127 macroeconomic variables and exclude principal components that have eigenvalues smaller than 2e-16. (If eigenvalues are too small, one cannot obtain the inverse matrix of the covariance matrix when generating $\boldsymbol{dz}_{v,t}$ in equation (3) later.).

Step 4: Using the final data set, perform a regression analysis.

Start with $\boldsymbol{v}_{t+1}$, a K-dimensional vector of state variables, i.e., five yield factors and macroeconomic variables at time t+1. The equation is as follows:

$$\nu_{t+1} = \gamma \cdot \nu_t + \alpha$$

From this, generate $\boldsymbol{dv_{t+1}}$ and $\boldsymbol{dz_{t,\,t}}$, which denotes the source of risk, using the following equations:

$d\nu_{t+1} = \nu_{t+1} - \left(\hat{\gamma} \cdot \nu_t + \hat{\alpha}\right)$ and $dz_{\nu,t} \equiv \Sigma_t^{-.5} d\nu_t \sim N(0, I)\sqrt{dt}$

Σ_t is the covariance matrix of $\boldsymbol{dv_t}$, and is equivalent to I (since $\boldsymbol{dv_t}$ is obtained as a result of conducting PCA on the time-series data).

Step 5: Collect the sigma and beta, the regression coefficients, using the following equations.

$\sigma^2_{f,n,t} \equiv std(df_{n,t})^2 \,/\, dt$ and $\vec{\beta}_{n,t} dt \equiv cov(dz_{\nu,t}, df_{n,t}) / \sigma^2_{f,n,t}$ ..

Step 6: Formulate a regression equation (1), which is rearranged as equation (2) as follows:

(1) $df_{n,t} \cdot \Delta = -\lambda_t' dz_{\nu,t} + \sigma^2_{f,n,t}\left(\Delta/2 + \lambda_t' \cdot \vec{\beta_{n,t}}\right)\Delta dt + \sigma_n \Delta dz_{n,t}$

(2) $df_{n,t} = \lambda_t' \cdot \left(-dz_{\nu,t}/\Delta + \sigma^2_{f,n,t} \cdot \vec{\beta_{n,t}} \cdot dt\right) + \sigma^2_{f,n,t} \cdot \Delta \cdot dt/2 + \epsilon$.

λ_t denotes the price of risks and is a linear function of risks, ν_t. In equation (2), $\sigma^2_{f,n,t} \cdot \Delta \cdot dt/2$ is a constant; therefore,

(3) $df_{n,t} = \alpha + \lambda_t' \cdot \left(-dz_{\nu,t}/\Delta + \sigma^2_{f,n,t} \cdot \vec{\beta_{n,t}} \cdot dt\right) + \epsilon$.

Step 7: Use equation (3) to obtain λ_t for each *t* in regression analysis. This completes an arbitrage-free affine term structure model with a large number of macro-financial factors.

3. Data

We use Liu and Wu's (2021) yield curve data set, which is also used by Bianchi, Buchner, and Tamoni (2020) who investigate bond risk premia using machine learning.[1] Using the data, we use regression and PCA methods to estimate up to 129 pricing factors, including five yield factors and 124 macroeconomic variables.

For macroeconomic variables, we collect monthly macroeconomic data from the Federal Reserve Economic Data (FRED) database.[2] We use *tcode* to transform the data before generating macroeconomic factors in reference to the method of McCracken and Ng (2015). Due to frequent missing values, more specifically, we use the five-step procedure to balance the unbalanced panel of macroeconomic variables (see Appendix II for details).

We start with the pool of 127 macroeconomic variables to collect macroeconomic data, following Ludvigson and Ng (2009). PCA is used to generate 127 principal components but three of the 127 factors have too small eigenvalues to find the inverse of the matrix; therefore, a final set of 124 factors is used for in-sample tests and a set of 121 factors for out-of-sample tests in addition to five yield factors. This makes the total number of factors 129 and 126 for in-sample and out-of-sample tests, respectively. The list of 124 macroeconomic variables is attached in Appendix III.

Table 1 describes summary statistics of the model-generated yields and observed yields. The sample period is from 1971:8 to 2021:2 for all the following figures and tables in this paper unless otherwise stated.

###Insert Table 1 about here###

The estimates in Table 1 show the fitting of the model-implied yields during the in-sample period. The results support that increasing the number of factors from five to 129 enhances the

[1] Source: https://sites.google.com/view/jingcynthiawu/yield-data
[2] Source: https://research.stlouisfed.org/econ/mccracken/fred-databases/

fitting of the yields. The statistics of the model-implied yields generated by using 129 factors are indeed nearly identical to those of the observed spot rates.

Finally, we employ a simple method to select relevant macroeconomic variables for the model. While PCA is widely used to verify the relative importance of all pricing factors, Lasso has the advantage of enabling the selection of only relevant factors while neglecting irrelevant ones. We exemplify how Lasso regression can be used to find, for example, a final set of 23 macroeconomic variables out of 124 macroeconomic variables in our test. Using this method in combination with our proposed model can simplify the construction process for predicting bond returns. A discussion of the results is included in subsection 4.3.

4. Empirical Tests

4.1. Yield fitting

To check the fitting of the forward rates implied by our model, we conduct time series and cross-sectional regressions for the period of 1972:8 to 2021:12. Figure 1 plots the time series yield fitting and estimates of term premia as well as predictability for one-month holding returns of the observed and model-implied yields.

###Insert Figure 1 about here###

All four graphs in Figure 1 plot the results generated by using all 129 factors, including five yield factors and 124 macroeconomic variables.[3] The upper two graphs show that the observed and model-implied yields are almost perfectly matched when measured over time. The bottom two graphs also show that the proposed model can predict one-month holding excess returns of the bonds almost perfectly while the term premia remain stable.

[3] Our model is used to generate four and five yield factors for comparison with prior affine models that use the same number of factors (e.g., CP and ACM models). The results are reported in Appendix IV.

###Insert Figure 2 about here###

Figure 2 plots cross-sectional regression results using the means and standard deviations of the observed and model-implied yields across different maturities ranging from 3 to 120 months. The graph on the left-hand side plots the unconditional means while the graph on the right-hand side plots the unconditional standard deviations of the observed and model-implied yields. The results show that the observed and model-implied yields are almost perfectly matched when tested cross-sectionally as well.[4]

4.2. Forecasting bond returns using yield and macroeconomic factors

Existing affine term structure model studies primarily focus on how a model generates yields that are close to observed yields. However, examining whether the proposed model can predict future returns during out-of-sample periods would also be equally important, especially for practitioners trying to generate excess returns using the model. Jang, Kang, and Lee (2020) examine the predictive power of the ACM model using Korean bond data. They use forward rates generated by the ACM model and empirically show that the difference between model-implied forward rates from the prior and present periods can be used to predict the difference between the actual forward rates from the prior and present periods. The tests are done for the in-sample and out-of-sample periods.

To explore whether our proposed model can be used to predict future returns using forward rates, similar to Jang et al. (2020), we use the following equation for regression:

$$\Delta f_{observed,\,t+1} = \alpha + \beta_t \left(f_{observed,t} - f_{model,t} \right) + \gamma_t \Delta f_{observed,\,t} + \epsilon_t$$

[4] Results generated from using four and five yield factors are reported in Appendix V.

To perform this regression, we convert the observed spot rates in our sample into forward rates because our model generates forward rates. The independent variable is the difference between the observed and model-implied forward rates at t that mature in n months, denoted as $\boldsymbol{\beta}$. The dependent variable is the difference between the forward rates at t and $t + 1$ that mature in n months. The difference between the forward rates at t-1 and t is used as a control variable and the coefficient is denoted, $\boldsymbol{\gamma}$.

###Insert Table 2 about here###

Table 2 reports the regression results generated by using our proposed model with four, five, and 129 factors, which include five yield factors and 124 macroeconomic variables. The coefficients and t-values for beta and gamma, i.e., a control variable, are reported. The results imply that the predictability of the model varies depending on the bonds' maturities but significantly increases as the number of factors increases from four to 129. For example, for the bonds with 24-month maturities, the model predicts the returns with far more significance when 129 factors are used (t-value of 2.947) than when 4 factors are used (t-value of 0.913). The same applies to the bonds with other maturities except for those with 48-month maturities where the significance of beta coefficients deteriorates as the number of factors increases.

###Insert Figure 3 about here###

Figure 3 plots only the t-values generated from the same regression, of which the results are reported in Table 2. The upper two graphs show that our model with four or five factors has a significant forecasting power for future bond returns using forward rates at t. However, when the number of factors increases to 129 as in the graphs in the second row, our model displays no forecasting ability as shown in the flat lines. We assume that this may be related to the small sample size, and therefore may be mitigated through data augmentation. Indeed, the predictive

power becomes significant with expanded data and the results are presented and discussed in the later section.

4.3. Selection of macroeconomic variables

Among the 129 factors used in the tests, 5 are yield factors and 124 are macroeconomic variables. This is a large set of factors that may be hard for practitioners to collect the data and run tests upon. Conducting PCA usually helps us examine the relative importance of all the variables; however, Lasso, which helps select only the relevant factors for empirical tests, may be more useful when we have a large number of factors. More specifically, Lasso regression can be used to minimize the number of factors, simplifying the model construction process and enabling economic interpretation of the outcomes generated by the model. In Table 2, we show how Lasso regression helps us narrow the number of factors for the empirical tests in our study.

###Insert Table 3 about here###

In this case, we choose only the variables that have *t*-values greater than 1.960. This leaves only 23 macroeconomic variables out of 121 macroeconomic variables collected from FRED. This is one example of how practitioners can use Lasso to minimize the number of macro factors to obtain principal components of their chosen number for the model. As the number of macroeconomic factors decreases, it would become easier to understand which macroeconomic variables specifically contribute to the model outcomes.

4.4. Model comparison

Yield fitting

To check the robustness of our model, we compare the yield-fitting results of our proposed model to those of the ACM model. Figure 4 plots different estimates, e.g., epsilons, standard deviations, and RMSEs, for comparison.

###Insert Figure 4 about here###

###Insert Table 4 about here###

Predictability

To check the robustness of our model in terms of predicting future bond returns using model-implied forward rates, we examine the predictive power of the prior affine models. The results are reported in Figure 5 and also in Table 5. As mentioned earlier, we use 121 macroeconomic variables, having eliminated three variables that have too small eigenvalues to make an inverse of the matrix. For out-of-sample tests, as a result, all 126 factors are used, including 5 yield factors and 121 macroeconomic factors.

###Insert Figure 5 about here###

In the bottom right graph, the RMSEs of all models range from zero to 0.005, i.e., the forecasting error is nearly zero across all maturities. This indicates that the forecasting power of the proposed model does not decay but even improves at some maturities. Some may argue that using only a few factors in the conventional affine model to predict future returns may be more convenient if the predictive power does not significantly improve; however, our proposed model can accommodate a large number of pricing factors, allowing for the use of any number of macroeconomic variables that affect bond pricing, without compromising the predictive ability.

##Insert Table 5 about here###

Table 5 reports the predictive power of alternative models, the ACM model with four and five factors. The results confirm the predictive power of the ACM model for future bond returns with

statistical significance across most maturities but is inferior to that of our proposed model with 129 factors shown in Table 2.

4.5. Data Augmentation

We perform additional tests to examine the effect of data augmentation on yield fitting. For earlier yield fitting, of which the results are reported in Figure 1 and Figure 2, the dependent variable is the difference between the observed and model-implied forward rates. The forward rates have n-month maturities in the multiples of three so that we have 40 data points in total (e.g. 3, 6, 9,...117, and 120-month).

To expand the data, we further divide the maturities into multiples of one, 0.25, and 0.125. For example, with the current data set, we do not have the data on forward rates at n=1, 2, 4, 5, …, 118, and 119. To expand the data to have monthly forward rates, we use a linear interpolation on forward rates at n=3 and 6 to obtain forward rates at n=4 and 5 and use a linear interpolation on forward rates at n=117 and 120 to obtain forward rates at n=118 and 119. The resulting data have 120 monthly forward rates in total instead of 40. When we further augment the data to have forward rates with n/4- and n/8-month maturities, we gain a total of 480 and 1,920 data points, respectively. Our proposed model's predictive power when measured with such expanded data is reported in Figure 6.

###Insert Figure 6 about here###

The most notable difference between Figure 3 and Figure 6 is that the model's prediction power when all 129 factors are used is significantly increased through data augmentation. The t-values stay between zero and one in Figure 3 whereas they become more significant and vary across different maturities when we have more data to measure the predictability of the model for

future bond returns. This would exemplify how data augmentation through linear interpolation, a commonly used technique in machine learning, can help mitigate the problem associated with a small sample size in asset pricing studies.

4.6. Model performance amid liquidity shock

We examine the proposed model's performance during the period of a liquidity shock. The purpose is to test whether the model generates superior performance amid market illiquidity. Figure 7 plots the RSMEs of the performance of our model and the ACM's during the COVID-19 pandemic, an out-of-sample period. The out-of-sample period is based on the reference date of 11 March 2020 when the COVID-19 pandemic was declared by the World Health Organization (WHO). The upper five lines represent the RMSEs during the in-sample period of 2008:3 to 2021:12 while the bottom five lines represent the RMSEs during the out-of-sample period of 2020:4 to 2021:12.

Insert Figure 7 about here

Figure 7 shows that the RSEMs of our model during the period of COVID-19 range from zero to 0.003 for all specifications reported in Figure 7. Such a result confirms that the predictability remains intact during the period of market distress caused by COVID-19, supporting the superior performance of our proposed model. The sample period in our study is as long as 50 years, which may strengthen generalize the conclusions; however, checking for the performance during a specific time period would further strengthen our conclusion.

4.7. Out-of-sample tests with Lasso

To check the robustness of the results, we test the model with different specifications in Lasso regressions. Figure 8 has figures plotting RMSEs to compare the performance of our proposed model to that of the ACM model with four and five factors. The model construction follows the same steps as those used to generate the results in Figure 3. The difference is that we use Lasso regression instead of OLS in one of the two locations, or both, as below:

$$\vec{\beta}_{n,t}dt \equiv cov(dz_{\nu,t}, df_{n,t})/\sigma^2_{f,n,t}\,.\; ie,\; regression\, coefficients \text{ and}$$

$$df_{n,t}\cdot\Delta = -\lambda_t' dz_{\nu,t} + \sigma^2_{f,n,t}\left(\Delta/2 + \lambda_t'\cdot\vec{\beta_{n,t}}\right)\Delta dt + \sigma_n \Delta dz_{n,t}.$$

Insert Figure 8 about here

In earlier tests, we use OLS to generate beta coefficients and lambda. In model type 1, we use Lasso in the first location and OLS in the second. In model type 2, we use Lasso instead of OLS in the second location only. In model type 3, we use Lasso in both the first and second locations.[5] Model type 0 is what we use to generate the main results using OLS in both locations. Using the same alpha within different ranges for model types 1, 2, and 3, we generate RMSEs for the out-of-sample tests. More specifically, we generate RMSEs using each estimate of alpha within the preset ranges for 12 maturities (i.e., 3, 6, 9, 12, 17, 24, 30, 36, 60, 72, 84, and 120 months) and add the generated RMSEs for each maturity.

The smallest calculated values are used to plot the graphs in Panel A of Figure 8. The results show that our proposed model performance is not compromised during the out-of-sample periods. The results in Panel B of Figure 8 also confirm that the predictability of our proposed

[5] In Lasso regression, L1 term (alpha) is a regularizing hyperparameter. When performing out-of-sample tests, estimates of alpha may differ depending on sample periods. Therefore, we would divide the data set into a train, validation, and test sets, and use the validation set for hyperparameter tuning. However, in this study, we use the same estimate of alpha for out-of-sample tests for simplicity. The estimate of alpha has a range of (0.0001, 0.001), (0.001, 0.01), and (0.001, 0.01) for model type 1, 2, and 3, respectively.

model is not compromised after data augmentation during the out-of-sample periods.

5. Conclusion

In this paper, we show that increasing the number of factors enhances the forecasting power of an affine model. We propose a new affine model that makes it possible and easy to use a large number of factors or even utilize machine learning methods in an affine model. In addition, we show that using the proposed methods involving PCA and Lasso can further simplify the process by enabling the users to choose a specific set of relevant macroeconomic variables. This saves the users from the wearing job of collecting exhaustive amounts of data or choosing ad-hoc variables without ground. Furthermore, we show that a small sample size problem can be mitigated from data augmentation through a machine learning method. The forecasting ability of the model is indeed enhanced in terms of its precision by expanding our data using Lasso.

In academia and practice, there is growing attention to the use of machine learning in asset pricing. Despite the advantages of using machine learning, such as generating outcomes with increased velocity and precision as well as less manpower, skeptics are often concerned about the possibility of data mining and breach of fundamental asset pricing assumptions such as the no-arbitrage condition (Bianchi et al., 2020; Chen et al., 2019). The proposed model in our study enables the use of machine learning with much flexibility while adhering to the assumption of no arbitrage. In that sense, our study may add some insight to the ongoing conversations about machine learning and asset pricing models, especially for those who value the fundamental assumptions of asset pricing but also recognize the importance of expanding the field by accepting the use of various machine learning methods.

In addition to investigating other machine learning methods in asset pricing, future studies can expand the literature by exploring the effect of behavioral traits of investors on asset pricing, for example, by using big data and machine learning to measure an investor's true risk appetite.

References

Adrian, T., Crump, R. K., & Moench, E. (2013). Pricing the term structure with linear regressions. *Journal of Financial Economics*, *110*(1), 110-138.

Ang, A., Hodrick, R. J., Xing, Y., & Zhang, X. (2006). The cross‐section of volatility and expected returns. *The Journal of Finance*, *61*(1), 259-299.

Bai, J., & Wang, P. (2016). Econometric analysis of large factor models. *Annual Review of Economics*, 8, 53-80.

Bansal, R., Khatchatrian, V., and Yaron, A., 2005, Interpretable asset markets?, *European Economic Review*, 49:531–60.

Bansal, R., & Yaron, A., 2004, Risks for the Long-Run: A Potential Resolution of Asset Pricing Puzzles, *Journal of Finance,* 59:1481–509.

Bianchi, D., Büchner, M., & Tamoni, A. (2021). Bond risk premiums with machine learning. The Review of Financial Studies, 34(2), 1046-1089.

Brandt, M. W., and Wang, K. Q., 2003, Time-varying risk aversion and unexpected inflation. *Journal of Monetary Economics,* 50:1457–98.

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

Chen, L., Pelger, M., & Zhu, J. (2019). Deep learning in asset pricing. *arXiv preprint arXiv:1904.00745*.

Chen, R., & Scott, L. (1993). Maximum Likelihood Estimation for a Multifactor Equilibrium Model of the Term Structure of Interest Rates. *The Journal of Fixed Income, 3 (3) 14-31.*

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

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

Duffie, D., & Kan, R. (1996). A yield‐factor model of interest rates. *Mathematical Finance*, 6(4), 379-406.

Estrella, A., & Hardouvelis, G. A., 1991, The term structure as a predictor of real economic activity. *The Journal of Finance*, 46(2), 555-576.

Fama, E. F. (1990). Stock returns, expected returns, and real activity. *The Journal of Finance*, *45*(4), 1089-1108.

Fama, E. F., & Bliss, R. R. (1987). The information in long-maturity forward rates. *The American Economic Review*, 680-692.

Giglio, Kelly, & Xiu. (2021). Factor Models, Machine Learning, and Asset Pricing. *Working paper*.

Israel, R., Kelly, B. T., & Moskowitz, T. J. (2020). Can Machines Learn Finance? *Journal of Investment Management*.

Harvey, C. R., Liu, Y., and Zhu, H., 2016, … and the cross-section of expected returns, *The Review of Financial Studies* 29(1), pp. 5-68.

Jang, G. Y., Kang, H. G., & Lee, D. J. (2021). An Extension of the Five‐factor Affine Term Structure Model: Predicting Future Bond Returns. *Asia‐Pacific Journal of Financial Studies*, *50*(6), 659-689.

Liu, Y., & Wu, J. C. (2021). Reconstructing the yield curve. Journal of Financial Economics, 142(3), 1395-1425.

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

Piazzesi, M. (2010). Affine term structure models. In *Handbook of financial econometrics: Tools and Techniques* (pp. 691-766). North-Holland.

Wachter, J., (2006). A consumption based model of the term structure of interest rates. *Journal of Financial Economics,* 79:365–99.

Wüthrich, K., & Zhu, Y. (2020). Omitted variable bias of Lasso-based inference methods: A finite sample analysis. *The Review of Economics and Statistics*, 1-47.

Figure 1. Time-series fit and term premia of model-implied yields using 129 factors.
This figure plots the yield fitting and term premium estimates, as well as predictability for one-month holding period excess returns of zero-coupon yield curve data for Treasuries with two- and ten-year maturities, as observed and implied by the ACM model and our proposed model with 129 pricing factors. Of the 120 factors, five are yield factors and 124 are macroeconomic variables. The sample data are constructed using Liu and Wu's (2021) yield curve data set. The sample period is from 1972:8 to 2021:12. For all graphs, solid lines represent observed yields and returns, dashed green lines represent model-implied yields and returns and dashed red lines represent the model-implied term premia.

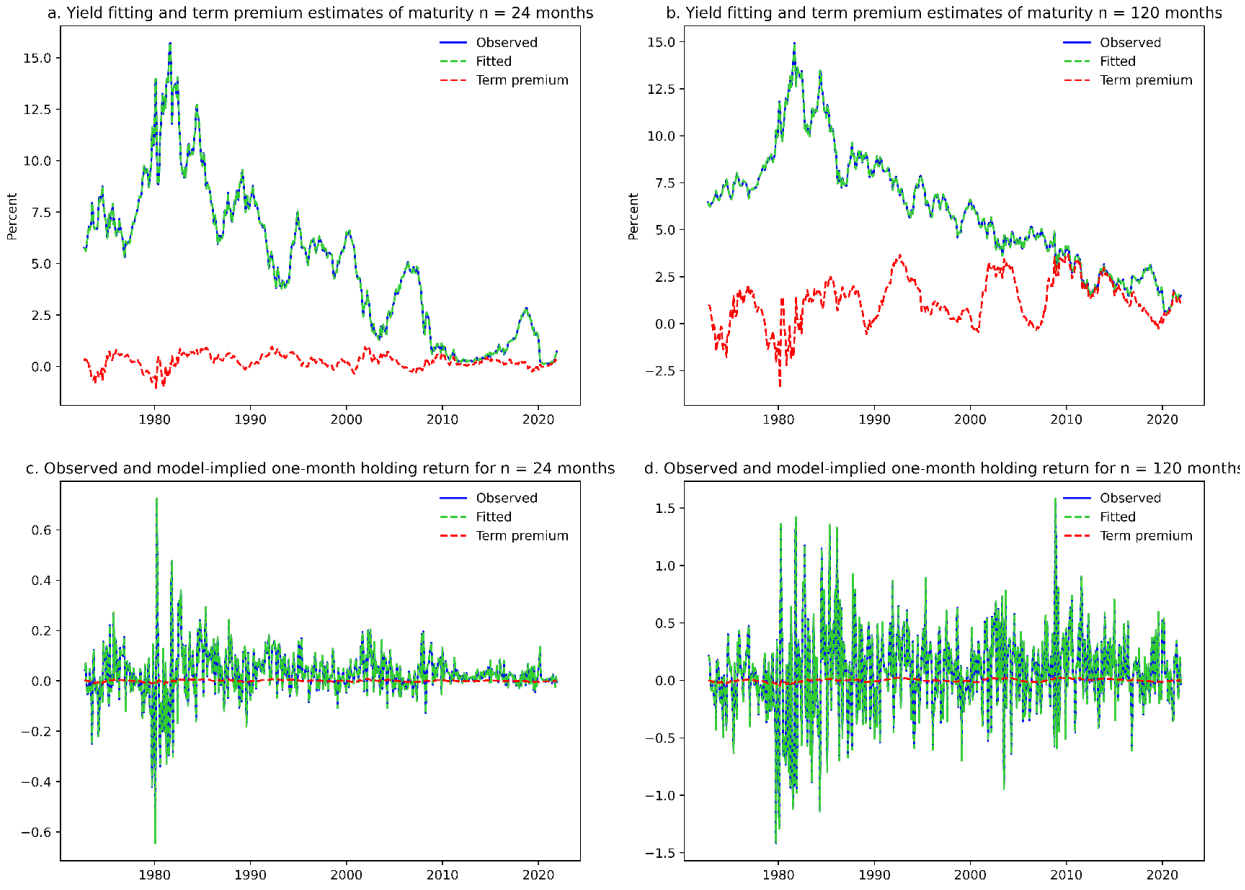


Figure 2. Cross-sectional fit of model-implied yields using 129 factors.
This figure plots the cross-sectional fit of the yields generated by using our proposed model with 129 pricing factors, including five yield factors and 124 macroeconomic variables. The sample data are constructed using Liu and Wu's (2021) yield curve data set. The sample period is from 1972:8 to 2021:12. The graph on the left-hand side plots the unconditional means while the graph on the right-hand side plots the unconditional standard deviations of the observed and model-implied yields.

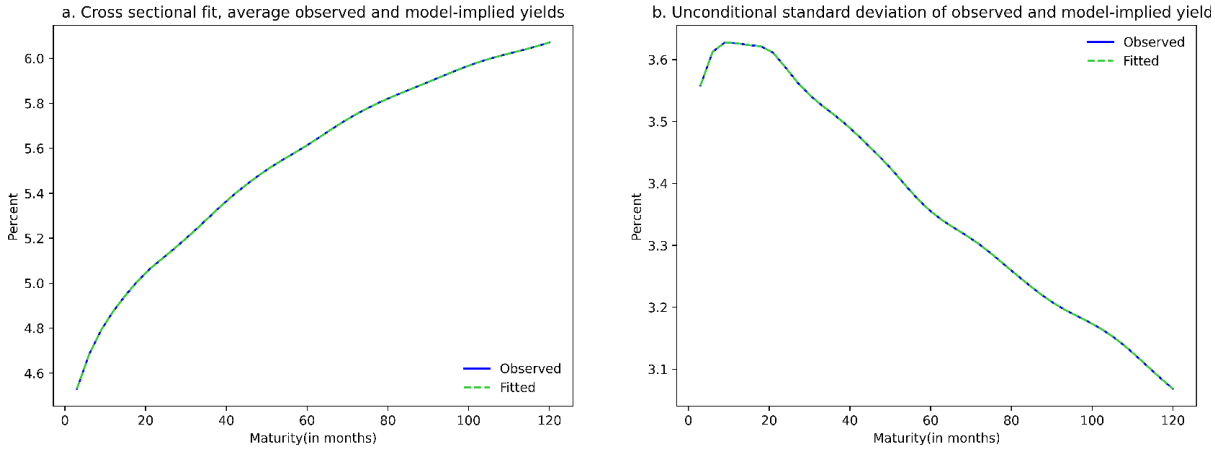


Figure 3. Predictive power across different maturities.

This figure plots the *t*-values generated from the regression used to test the predictive power of our proposed model. The following equations are used:

1) without control variables

$$\Delta f_{observed,\,t+1} = \alpha + \beta_t\left(f_{observed,\,t} - f_{model,\,t}\right) + \epsilon_t$$

2) with control variables

$$\Delta f_{observed,\,t+1} = \alpha + \beta_t\left(\ f_{observed,t} - f_{model,t}\right) + \gamma_t \Delta f_{observed,\,t} + \epsilon_t$$

The independent variable is the difference between the observed and model-implied forward rates at t that mature in n months, denoted as **β**. The dependent variable is the difference between the forward rates at t and $t+1$ that mature in n months. The difference between the forward rates at t-1 and t is used as a control variable and the coefficient is denoted, **γ**. The sample data are constructed using Liu and Wu's (2021) yield curve data set. The sample period is from 1972:8 to 2021:12.

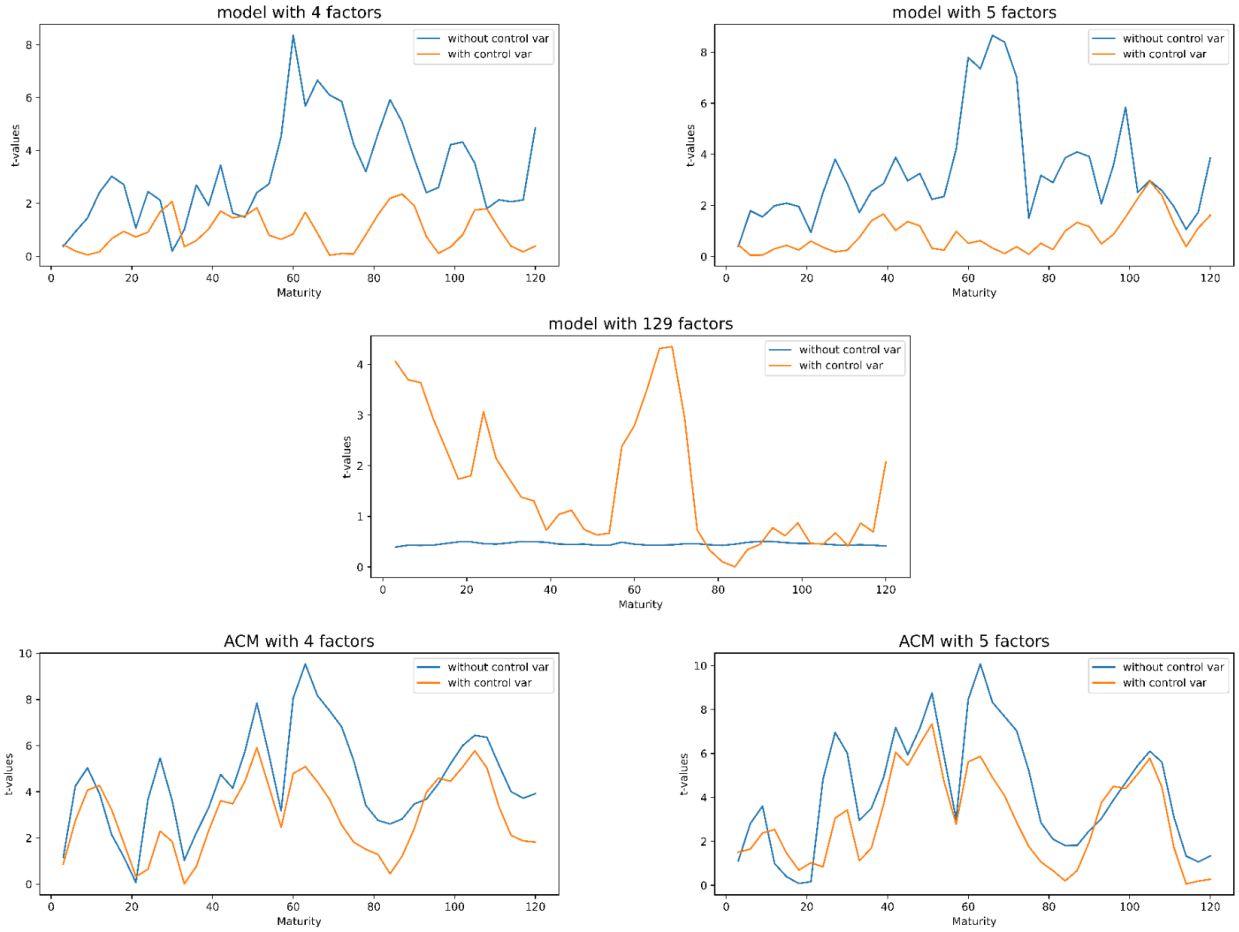


Figure 4. In-sample model performance comparison.

The figure plots in-sample model performance for comparison using two metrics such as the percentage valuation errors of the predicted value of $dfwd$ rate (ϵ) and root mean squared error (RMSE). The percentage valuation errors measure the accuracy in predicting dfwd rate and is defined as $\epsilon \equiv \widehat{dfwd} \,/\, dfwd \,-\, 1$ where $dfwd$ is the difference of forward rate and $\widehat{dfwd}$ is the corresponding model estimate. RMSE measures the difference between the actual value and predicted value, and it is defined as $RMSE(\widehat{dfwd}) \equiv \sqrt{\frac{1}{N}\sum (dfwd \,-\, \widehat{dfwd})^2}$.

Figure 5. Out-of-sample model performance comparison.

We divide the data into a training set and a test set in the ratio of 7:3. The sample period for the training set is from 1972:8 to 2007:2 while that for the test set is from 2007:3 to 2021:12. We estimate prediction values, denoted as $\widehat{dfwd}$, for training set periods in the following steps:

Step 1. Use the data from 1972:8 to 2007:2 for yield fitting

Step 2. Calculate $\widehat{dfwd}$ by using model-implied and observed forward rates in the following equation:

$$\Delta f_{observed,\,t+1} = \alpha + \beta_t \left(f_{observed,t} - f_{model,t} \right) + \gamma_t \Delta f_{observed,\,t} + \epsilon_t$$

Step 3. Expand the period of training set to 1972:2 to 2007:3 and repeat Step 1 and 2.

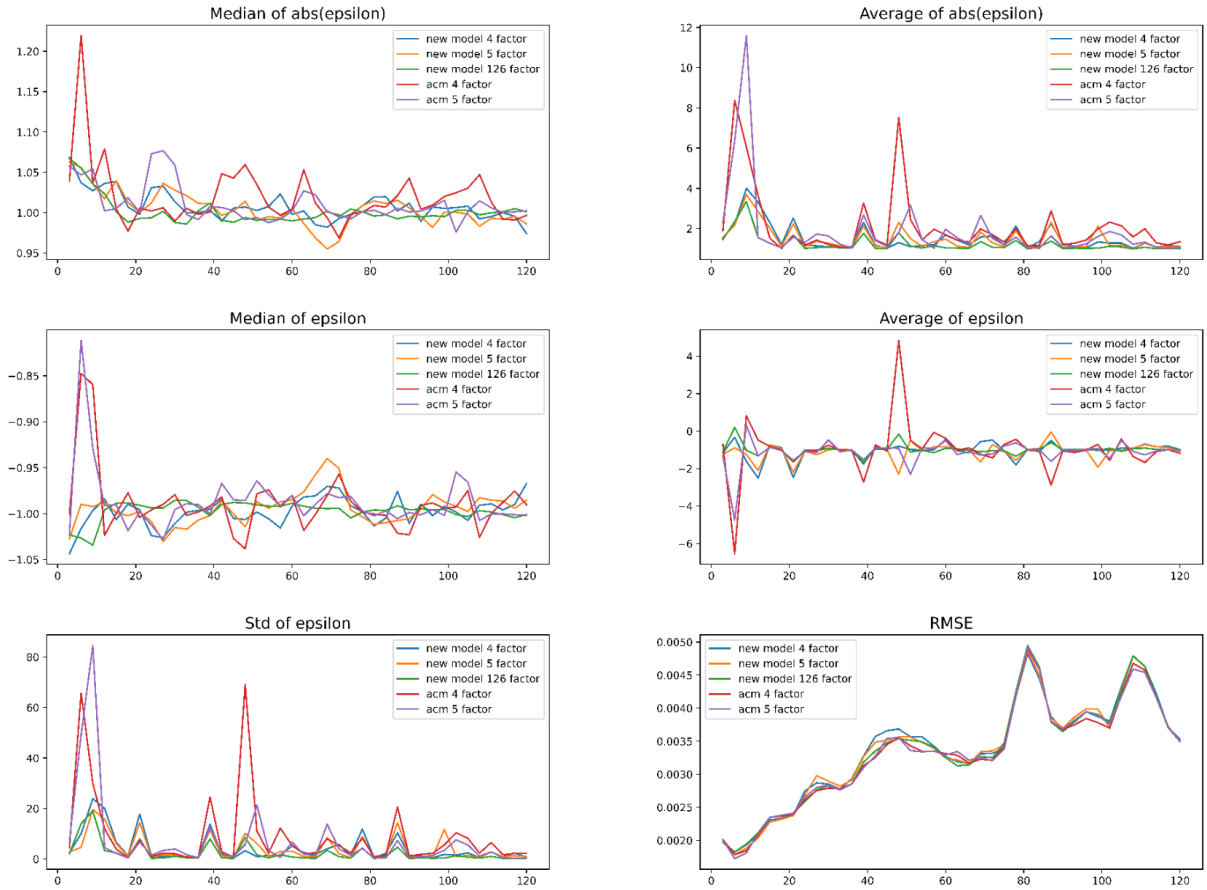


Figure 6. Predictive power of the model using data augmentation.
This figure plots the *t*-values of beta and gamma coefficients generated from the regressions to test the predictive power of the model using expanded data. The following equations are used:

1) without control variables

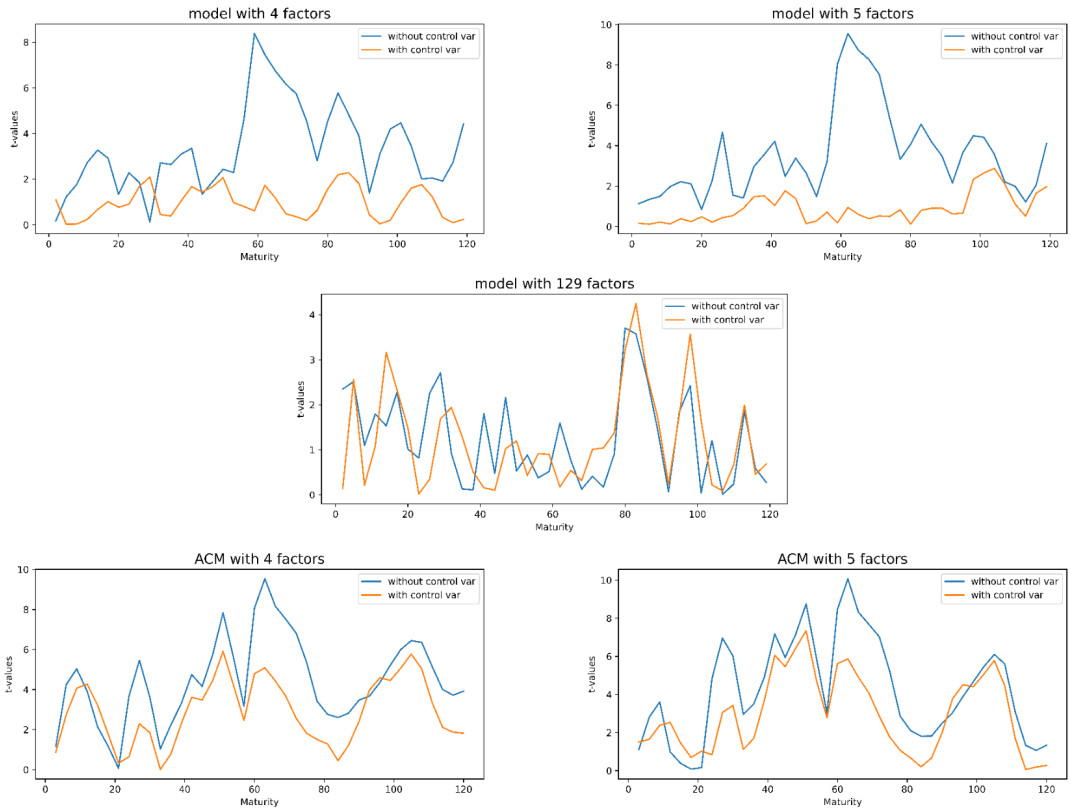
$$\Delta f_{observed,\,t+1} = \alpha + \beta_t (f_{observed,\,t} - f_{model,\,t}) + \epsilon_t$$

2) with control variables

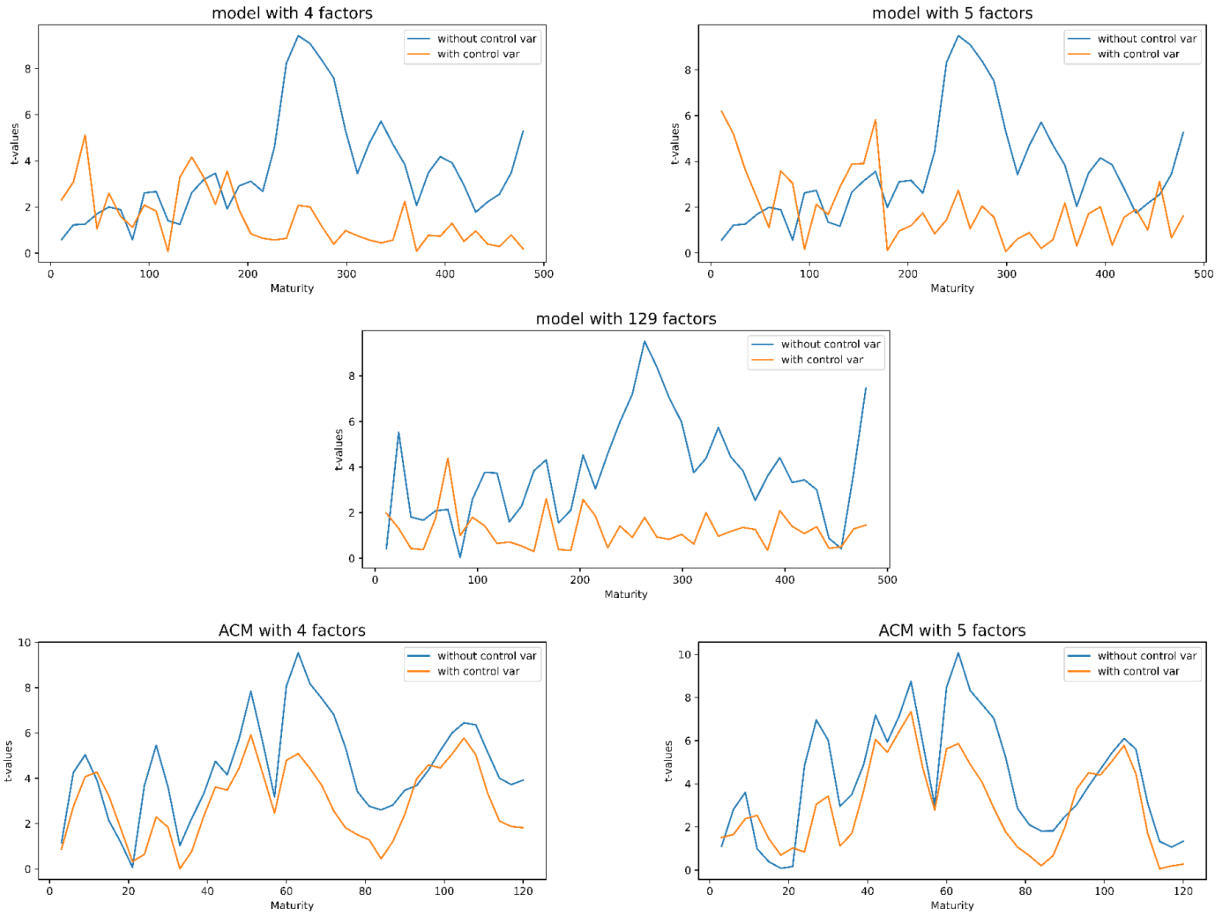
$$\Delta f_{observed,\,t+1} = \alpha + \beta_t \left(f_{observed,t} - f_{model,t} \right) + \gamma_t \Delta f_{observed,\,t} + \epsilon_t$$

The sample data in this study has a total of 40 data points, consisting of forward rates with *n*-month maturities that are the multiples of 3-month. Through linear interpolation, we augment the data to have 120, 480, and 1,920 data points that are forward rates with 1/*n*-, 0.25/*n*-, and 0.125/*n*-month maturities, respectively. Panel A plots the results from having the data of 120 forward rates implied by our proposed model with four, five, and 129 factors, which includes 5 yield factors and 124 macroeconomic variables. Panel B and Panel C plot the results from having the data of 480 and 1,920 forward rates, implied by the model with four, five, and 129 factors.

Panel A. *t*-values of beta and gamma coefficients when using 120 data points.



Panel B. *t*-values of beta and gamma coefficients when using 480 data points.



Panel C. *t*-values of beta and gamma coefficients when using 1,920 data points.

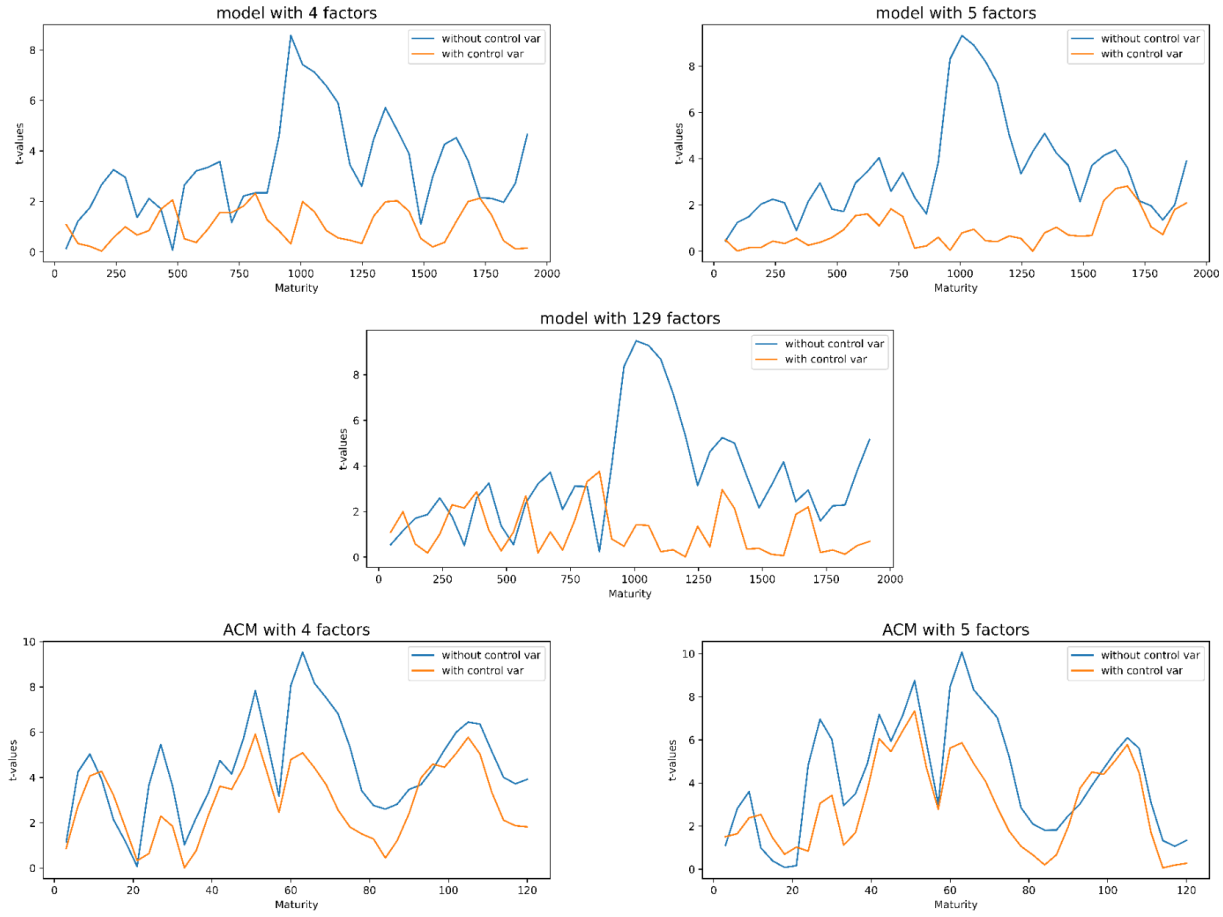


Figure 7. Out-of-sample model performance comparison during COVID-19.
This figure presents the value of root mean squared error (RMSE) during the COVID-19 pandemic, an out-of-sample period. The out-of-sample period is based on the reference date of 11 March 2020 when the COVID-19 pandemic was declared by the World Health Organization (WHO). The sample data are constructed using Liu and Wu (2021)'s yield curve data set. The upper five lines represent the RMSEs during the in-sample period of 2008:3 to 2021:12 while the bottom five lines represent the RMSEs during the out-of-sample period of 2020:4 to 2021:12.

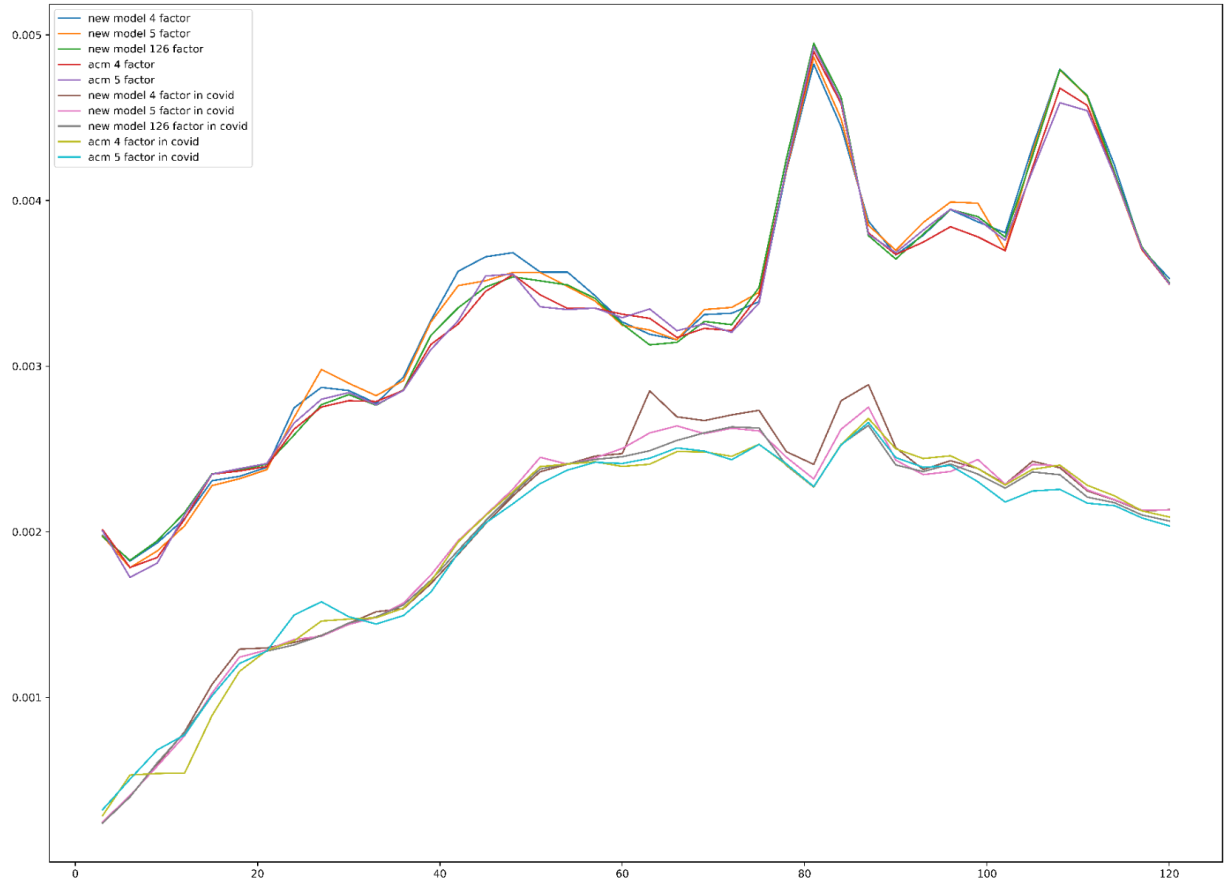


Figure 8. Out-of-sample model performance comparison using Lasso.

This figure shows the predictability of the ACM and our proposed model during the out-of-sample period using Lasso. Panel A plots the RMSEs using the original dataset of 40 maturities in the ACM model with four and five factors as well as different model specifications of Lasso. Panel B plots the RMSEs using the augmented data, consisting of 120, 480, adn 1,920 data points, in the ACM with four and five factors and our proposed model with four, five, and 129 factors. 129 factors include five yield factors as well as 124 macroeconomic variables. For performance tests during the out-of-sample periods, we divide the data into a training set and a test set in the ratio of 7:3. The sample period for the training set is from 1972:8 to 2007:2 while that for the test set is from 2007:3 to 2021:12. We estimate prediction values, denoted as $\widehat{dfwd}$, for training set periods in the following steps:

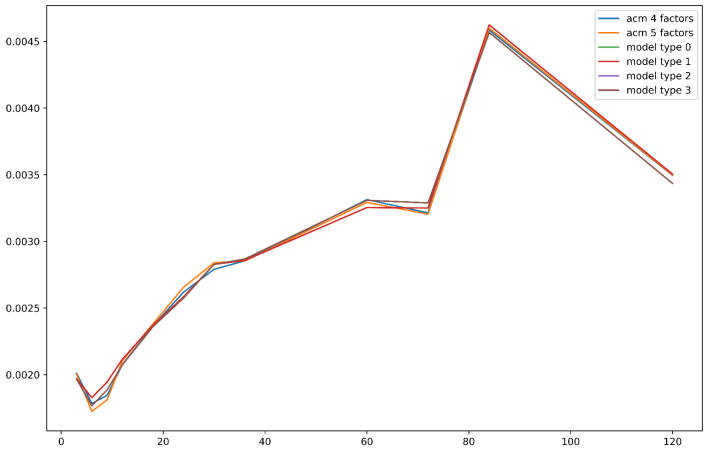
Step 1. Use the data from 1972:8 to 2007:2 for yield fitting.

Step 2. Calculate $\widehat{dfwd}$ by using model-implied and observed forward rates in the following equation:

$$\Delta f_{observed,\, t+1} = \alpha + \beta_t \left(f_{observed,t} - f_{model,t} \right) + \gamma_t \Delta f_{observed,\, t} + \epsilon_t$$

Step 3. Expand the period of training set to 1972:2 to 2007:3 and repeat Step 1 and 2.

Panel A. With the original data set of 40 maturities.



Panel B. With data augmentation.

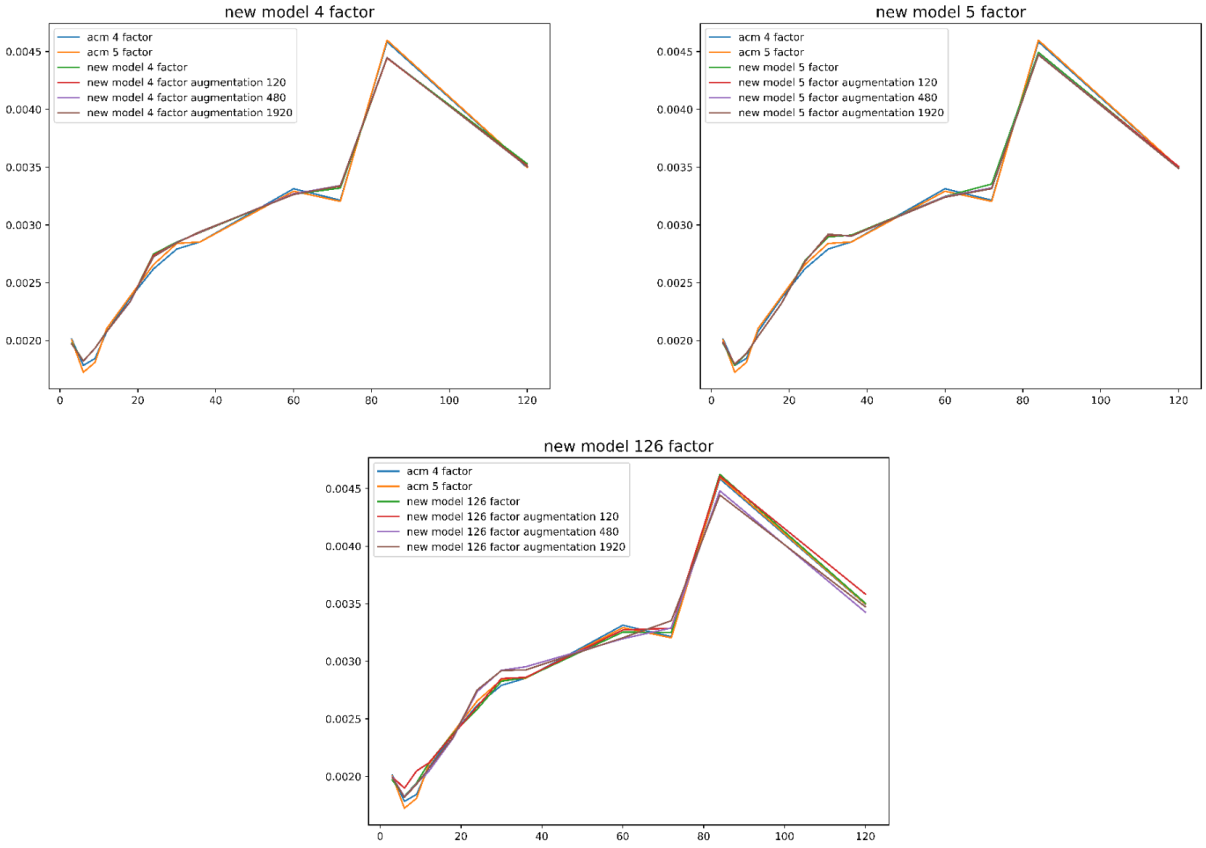


Table 1. Summary statistics of observed and model-implied yields.

This table reports the summary statistics of the observed and model-implied yields from the period of 1971:8 to 2021:2. Panel A shows the summary statistics of the spot rates observed for the same period. Panel B reports the summary statistics of the spot rates generated by using five yield factors in our proposed model. Five factors include yield, slope, curvature, and two additional higher-order factors as generated by Adrian, Crump, and Moench (2013). Panel C reports the summary statistics of the spot rates generated by using five yield factors in addition to 124 macroeconomic variables in our proposed model. The fitted forward rates generated by the model are transformed to spot rates, which are used for the summary statistics to compare with the observed spot rates. For all panels, the number of observations (*count*), average values (*mean*), standard deviations (*std*), minimum values (*min*), 25% (*25%),* 50% (*50%),* and 75% (*75%)* percentile values, and maximum values (*max*) are reported.

Panel A. Observed yields

maturities	3	6	9	12	18	24	30	36	60	72	84	120
count	592	592	592	592	592	592	592	592	592	592	592	592
mean	0.045	0.047	0.048	0.049	0.050	0.051	0.052	0.053	0.056	0.058	0.059	0.061
std	0.036	0.036	0.036	0.036	0.036	0.036	0.035	0.035	0.034	0.033	0.032	0.031
min	0.000	0.000	0.000	0.001	0.001	0.001	0.001	0.001	0.002	0.003	0.004	0.005
25%	0.012	0.012	0.014	0.015	0.015	0.016	0.018	0.020	0.025	0.028	0.030	0.037
50%	0.049	0.050	0.051	0.052	0.053	0.054	0.055	0.055	0.058	0.058	0.060	0.061
75%	0.066	0.069	0.071	0.071	0.073	0.074	0.075	0.076	0.077	0.078	0.079	0.079
max	0.159	0.161	0.161	0.160	0.159	0.157	0.155	0.156	0.152	0.150	0.150	0.149

Panel B. Model-implied yields using 5 factors

maturities	3	6	9	12	18	24	30	36	60	72	84	120
count	592	592	592	592	592	592	592	592	592	592	592	592
mean	0.045	0.047	0.048	0.049	0.050	0.051	0.052	0.053	0.056	0.057	0.059	0.061
std	0.036	0.036	0.036	0.036	0.036	0.036	0.035	0.035	0.034	0.033	0.032	0.031
min	- 0.008	- 0.006	- 0.004	- 0.003	- 0.003	- 0.002	- 0.002	- 0.000	0.002	0.003	0.004	0.005
25%	0.012	0.013	0.014	0.015	0.016	0.017	0.018	0.020	0.025	0.028	0.031	0.037
50%	0.048	0.050	0.051	0.052	0.054	0.054	0.055	0.056	0.058	0.059	0.059	0.061
75%	0.066	0.069	0.071	0.072	0.073	0.074	0.075	0.076	0.077	0.078	0.079	0.079
max	0.159	0.163	0.164	0.163	0.164	0.161	0.159	0.160	0.156	0.154	0.153	0.149

Panel C. Model-implied yields using 129 factors

maturities	3	6	9	12	18	24	30	36	60	72	84	120
count	592	592	592	592	592	592	592	592	592	592	592	592
mean	0.045	0.047	0.048	0.049	0.050	0.051	0.052	0.053	0.056	0.058	0.059	0.061
std	0.036	0.036	0.036	0.036	0.036	0.036	0.035	0.035	0.034	0.033	0.032	0.031
min	0.000	0.000	0.000	0.001	0.001	0.001	0.001	0.001	0.002	0.003	0.004	0.005
25%	0.012	0.012	0.014	0.015	0.015	0.016	0.018	0.020	0.025	0.028	0.030	0.037
50%	0.049	0.050	0.051	0.052	0.053	0.054	0.055	0.055	0.058	0.058	0.060	0.061
75%	0.066	0.069	0.071	0.071	0.073	0.074	0.075	0.076	0.077	0.078	0.079	0.079
max	0.159	0.161	0.161	0.160	0.159	0.157	0.155	0.156	0.152	0.150	0.150	0.149

Table 2. Predictive power of the proposed model.
This table presents the regression results for the predictive power of our proposed model for future bond returns using the following equation:

$$\Delta f_{observed,\,t+1} = \alpha + \beta_t \left(f_{observed,t} - f_{model,t} \right) + \gamma_t \Delta f_{observed,\,t} + \epsilon_t$$

The independent variable is the difference between the observed and model-implied forward rates at t that mature in n months, denoted as **β**. The dependent variable is the difference between the forward rates at t and $t + 1$ that mature in n months. The difference between the forward rates at t-1 and t is used as a control variable and the coefficient is denoted, **γ**. The sample data are constructed using the Liu and Wu's(2021) yield curve data set. The sample period is from 1972:8 to 2021:12.

	Four factors					Five factors					129 factors				
matuirties	constant	beta	t-value	gamma	t-value	constant	beta	t-value	gamma	t-value	constant	beta	t-value	gamma	t-value
3	- 0.000	- 0.107	- 0.441	- 0.106	- 0.406	- 0.000	- 0.150	- 0.458	- 0.099	- 0.370	- 0.000	- 0.000	- 3.897	- 0.210	- 1.777
6	- 0.000	- 0.049	- 0.200	0.142	1.233	- 0.000	0.013	0.051	0.142	1.217	- 0.000	0.000	3.907	0.141	1.230
12	- 0.000	0.042	0.169	- 0.118	- 1.006	- 0.000	- 0.071	- 0.302	- 0.122	- 1.034	- 0.000	0.000	2.952	- 0.117	- 1.003
24	- 0.000	0.196	0.913	- 0.050	- 0.488	- 0.000	- 0.073	- 0.360	- 0.053	- 0.516	- 0.000	0.000	2.947	- 0.048	- 0.473
48	- 0.000	0.288	1.532	- 0.063	- 0.601	- 0.000	- 0.233	- 1.197	- 0.104	- 0.978	- 0.000	0.000	0.724	- 0.078	- 0.748
60	- 0.000	0.150	0.849	- 0.067	- 0.683	- 0.000	- 0.099	- 0.519	- 0.073	- 0.745	- 0.000	0.000	2.813	- 0.071	- 0.725
72	- 0.000	0.017	0.101	0.078	0.808	- 0.000	0.077	0.384	0.076	0.784	- 0.000	- 0.000	- 2.913	0.077	0.803
96	- 0.000	0.067	0.114	0.029	0.262	- 0.000	0.188	0.868	0.024	0.215	- 0.000	0.000	1.245	0.028	0.258
120	- 0.000	- 0.069	- 0.390	0.330	2.835	- 0.000	0.412	1.619	0.332	2.857	- 0.000	- 0.000	- 0.192	0.329	2.830

Table 3. Relative importance of macroeconomic variables.

This table reports the relative importance of macroeconomic variables using Lasso. We conduct Lasso regressions with different specifications in model type 1, 2, and 3 (*see* the description of Figure 8). The results are nearly identical regardless of the model type, so we randomly choose model type 2 and the best alpha generated from the out-of-sample test to obtain coefficients for the five yield factors and 121 macroeconomic variables. With eigenvalues obtained from conducting PCA on the data of 127 macroeconomic variables, we calculate a lasso coefficient multiplied by the eigenvector for each macroeconomic variable, of which the average and standard deviation are used to calculate a *t*-value of each macroeconomic variable using the general equation, *t*-value = mean / std/$\sqrt{N}$.

No.	FRED	Category	Description	t-value
1	IPDMAT	Output and Income	IP: Durable Materials	- 1.988
2	CES1021000001	Labor Market	All Employees: Mining and Logging: Mining	- 2.716
3	USWTRADE	Labor Market	All Employees: Wholesale Trade	2.129
4	AWHMAN	Labor Market	Avg Weekly Hours : Manufacturing	2.216
5	HOUST	Consumption and Orders	Housing Starts: Total New Privately Owned	2.879
6	HOUSTW	Consumption and Orders	Housing Starts, West	3.161
7	PERMIT	Consumption and Orders	New Private Housing Permits (SAAR)	2.721
8	PERMITMW	Consumption and Orders	New Private Housing Permits, Midwest (SAAR)	2.719
9	PERMITW	Consumption and Orders	New Private Housing Permits, West (SAAR)	2.465
10	ACOGNO	Orders and Inventories	New Orders for Consumer Goods	- 2.974
11	ISRATIOx	Orders and Inventories	Total Business: Inventories to Sales Ratio	2.056
12	REALLN	Money and Credit	Real Estate Loans at All Commercial Banks	2.102
13	S&P 500	Stock Market	S&P's Common Stock Price Index: Composite	- 2.017
14	S&P PE ratio	Stock Market	S&P's Composite Common Stock: Price-Earnings Ratio	- 2.029
15	FEDFUNDS	Interest rate and Exchange Rates	Effective Federal Funds Rate	2.394
16	GS1	Interest rate and Exchange Rates	1-Year Treasury Rate	2.558
17	TB6SMFFM	Interest rate and Exchange Rates	6-Month Treasury C Minus FEDFUNDS	- 2.036
18	AAAFFM	Interest rate and Exchange Rates	Moody's Aaa Corporate Bond Minus FEDFUNDS	- 2.615
19	EXSZUSx	Interest rate and Exchange Rates	Switzerland / U.S. Foreign Exchange Rate	2.320
20	WPSFD49502	Prices	Producer Price Index by Commodity: Final Demand: Personal Consumption Goods	- 2.648
21	CUSR0000SAC	Prices	CPI : Commodities	- 2.796
22	CUSR0000SAS	Prices	CPI : Services	- 2.002
23	PCEPI	Prices	Personal Cons. Expend: Chain Index	- 2.703

Table 4. Summary statistics of model-implied rates using alternative models.

This table reports the summary statistics of the model-implied yields using alternative models for the period of 1971:8 to 2021:2. Panel A reports the summary statistics of the fitted spot rates generated by using four yield factors in the model created by Adrian, Crump, and Moench (2013). Four factors include yield, slope, curvature, and CP factor generated by the ACM model. Panel B reports the summary statistics of the fitted spot rates generated by using five yield factors in the ACM model. Five factors include yield, slope, curvature, and two additional higher-order factors as generated by the ACM model. For all panels, the number of observations (*count*), average values (*mean*), standard deviations (*std*), minimum values (*min*), 25% (*25%),* 50% (*50%),* and 75% (*75%)* percentile values, and maximum values (*max*) are reported.

Panel A. ACM model with 4 factors

	3	6	9	12	18	24	30	36	60	72	84	120
count	592	592	592	592	592	592	592	592	592	592	592	592
mean	0.045	0.046	0.047	0.048	0.050	0.051	0.052	0.053	0.056	0.057	0.058	0.060
std	0.036	0.036	0.036	0.036	0.036	0.036	0.035	0.035	0.034	0.033	0.032	0.030
min	- 0.001	- 0.000	0.000	0.000	0.000	0.001	0.001	0.002	0.002	0.003	0.004	0.006
25%	0.012	0.012	0.014	0.014	0.016	0.017	0.018	0.020	0.025	0.028	0.030	0.037
50%	0.049	0.049	0.050	0.051	0.053	0.054	0.055	0.056	0.058	0.058	0.059	0.060
75%	0.067	0.068	0.070	0.070	0.072	0.074	0.075	0.076	0.077	0.078	0.078	0.079
max	0.160	0.160	0.159	0.158	0.157	0.156	0.155	0.155	0.153	0.151	0.149	0.145

Panel B. ACM model with 5 factors

	3	6	9	12	18	24	30	36	60	72	84	120
count	592	592	592	592	592	592	592	592	592	592	592	592
mean	0.045	0.047	0.048	0.049	0.050	0.051	0.052	0.053	0.056	0.057	0.058	0.061
std	0.036	0.036	0.036	0.036	0.036	0.036	0.036	0.035	0.034	0.033	0.032	0.031
min	- 0.000	0.000	0.000	0.001	0.001	0.001	0.001	0.001	0.002	0.003	0.004	0.006
25%	0.012	0.013	0.013	0.014	0.015	0.016	0.018	0.020	0.025	0.028	0.030	0.037
50%	0.049	0.049	0.050	0.051	0.053	0.054	0.055	0.056	0.058	0.058	0.059	0.061
75%	0.066	0.069	0.071	0.071	0.073	0.074	0.075	0.076	0.077	0.078	0.078	0.079
max	0.159	0.160	0.160	0.160	0.158	0.157	0.157	0.156	0.152	0.150	0.149	0.149

Table 5. Predictive power of alternative models.
This table presents the regression results for the predictive power of the ACM model with four and four and five factors for future bond returns using the following equation:

$$\Delta f_{observed,\,t+1} = \alpha + \beta_t \left(f_{observed,t} - f_{model,t} \right) + \gamma_t \Delta f_{observed,\,t} + \epsilon_t$$

The independent variable is the difference between the observed and model-implied forward rates at t that mature in n months, denoted as **β**. The dependent variable is the difference between the forward rates at t and $t + 1$ that mature in n months. The difference between the forward rates at t-1 and t is used as a control variable and the coefficient is denoted, **γ**. The sample data are constructed using Liu and Wu's (2021) yield curve data set. The sample period is from 1972:8 to 2021:12.

	ACM model with 4 factors					ACM model with 5 factors				
maturities	constant	beta	t-value	gamma	t-value	constant	beta	t-value	gamma	t-value
3	- 0.000	- 0.136	- 0.870	- 0.183	- 1.503	- 0.000	- 0.817	- 1.511	- 0.151	- 1.219
6	0.000	- 0.365	- 2.746	0.064	0.545	0.000	- 0.331	- 1.652	0.091	0.764
12	0.000	- 0.463	- 4.276	- 0.148	- 1.284	0.000	- 0.662	- 2.538	- 0.077	- 0.655
24	- 0.000	- 0.067	- 0.641	- 0.043	- 0.416	- 0.000	- 0.096	- 0.843	- 0.043	- 0.416
42	0.000	- 0.293	- 3.612	0.051	0.473	0.000	- 0.626	- 6.060	- 0.006	- 0.060
60	0.000	- 0.270	- 4.791	- 0.018	- 0.192	- 0.000	- 0.370	- 5.622	- 0.010	- 0.103
72	0.000	- 0.122	- 2.560	0.060	0.629	0.000	- 0.158	- 2.868	0.061	0.636
96	0.000	- 0.145	- 4.589	0.016	0.148	- 0.000	- 0.155	- 4.510	0.033	0.308
120	- 0.000	- 0.039	- 1.819	0.328	2.826	- 0.000	0.007	0.277	0.328	2.821

Appendix I. The Proposed Affine Model Construction.

Let us specify forward rates as:

$$F_{n,t} \equiv F(t,n,n+\Delta) = \exp\left(-f_{n,t}\cdot\Delta\right)$$
$$F_{n,t+dt} \equiv F(t+dt,n,n+\Delta) = \exp\left(-f_{n,t+dt}\cdot\Delta\right)$$

$F(t,n,n+\Delta)$ is the value of the forward contract at n, but determined at t $(n>t)$ that pays one dollar at $n+\Delta$. Therefore, the following relationship holds:

$$F(t,n,n+\Delta) = Z(t,n+\Delta)/Z(t,n)$$

$Z(t,n)$ is the value of a zero-coupon bond at t that pays one dollar at n $(n>t)$.

In addition, our affine pricing kernel is assumed to be:

$$M_{t+dt} = \exp\left(-r_t dt - \lambda_t'\lambda_t/2\cdot dt - \lambda_t' dz_{\nu,t}\right)$$
$$dz_{\nu,t} \equiv \Sigma_t^{-.5} d\nu_t \sim N(0,I)\sqrt{dt}$$
$$\lambda_t \equiv (\lambda_0 + \lambda_1\nu_t)\Delta$$

if $d\nu_t$ is from PCA of time-series data, Σ_t is I.
$dz_{\nu,t}$ is the source of risks. Therefore, λ_t denotes the price of the risks.

Our model is called 'affine' because the price of risk (λ_t) is a linear function of risks (ν_t).
ν_t is a *K*-dimensional vector of state variables (eg, macro-financial variables, big data).

The value of entering two offsetting forward contracts is zero, ie, $E(M\cdot dF) = 0$. Hence, the definition of pricing kernel implies:

$$E(M_{t+dt}\cdot dF_{n,t}/F_{n,t}) = 0$$

$$\iff E\left(M_{t+dt}\cdot\frac{\exp\left(-f_{n,t+dt}\cdot\Delta\right)-\exp\left(-f_{n,t}\cdot\Delta\right)}{\exp\left(-f_{n,t}\cdot\Delta\right)}\right) = 0$$

$$\iff E(M_{t+dt}\cdot[\exp\left(-\Delta\cdot(f_{n,t+dt}-f_{n,t})\right)-1]) = 0$$

$$\iff E\left(M_{t+dt}\cdot\left(e^{-df_{n,t}\cdot\Delta}-1\right)\right) = 0$$

$$\iff E\Big(M_{t+dt}\cdot e^{-df_{n,t}\cdot\Delta}\Big) = E(M_{t+dt}\cdot 1) = e^{-r_t\cdot dt}$$

$$\iff E\Big[\exp\left(-\lambda_t'\lambda_t/2\cdot dt - \lambda_t' dz_{\nu,t}\right)\cdot e^{-r_t\cdot dt}\cdot e^{-df_{n,t}\cdot\Delta}\Big] = e^{-r_t\cdot dt}$$

$$\iff E\big(\exp\left(-\lambda_t'\lambda_t/2\cdot dt - \lambda_t' dz_{\nu,t} - df_{n,t}\Delta\right)\big) = 1$$

Applying Ito's lemma produces:

$$1 = E\Big(1 - \lambda_t'\lambda_t/2\cdot dt - \lambda_t' dz_{\nu,t} - df_{n,t}\Delta + \left(\lambda_t'\lambda_t/2\cdot dt + \lambda_t' dz_{\nu,t} + df_{n,t}\Delta\right)^2/2\Big)$$

$$\iff E(df_{n,t}\cdot\Delta) = E\Big(-\lambda_t'\lambda_t/2\cdot dt - \lambda_t' dz_{\nu,t} + \left(\lambda_t'\lambda_t/2\cdot dt + \lambda_t' dz_{\nu,t} + df_{n,t}\Delta\right)^2/2\Big)$$

$$\iff E(df_{n,t}\cdot\Delta) = E\Big(-\lambda_t' dz_{\nu,t} + (df_{n,t})^2\Delta^2/2 + \lambda_t' dz_{\nu,t}\cdot df_{n,t}\Delta\Big)$$

To use notations as follows:

$$\sigma_{f,n,t}^2 dt \equiv (df_{n,t})^2$$
$$\vec{\beta}_{n,t} dt \equiv cov(dz_{\nu,t}, df_{n,t})/\sigma_{f,n,t}^2 \,.\, ie,\ regression\, coefficients$$

$$\implies E(df_{n,t}\cdot\Delta) = E\Big(-\lambda_t' dz_{\nu,t} + \sigma_{f,n,t}^2\cdot dt\Delta^2/2 + \lambda_t'\cdot\vec{\beta_{n,t}}\cdot\sigma_{f,n,t}^2\Delta dt\Big)$$
$$\iff E(df_{n,t}\cdot\Delta) = E\Big[-\lambda_t' dz_{\nu,t} + \sigma_{f,n,t}^2\Big(\Delta/2 + \lambda_t'\cdot\vec{\beta_{n,t}}\Big)\Delta dt\Big]$$

Then, the empirical design becomes:

$$df_{n,t}\cdot\Delta = -\lambda_t' dz_{\nu,t} + \sigma_{f,n,t}^2\Big(\Delta/2 + \lambda_t'\cdot\vec{\beta_{n,t}}\Big)\Delta dt + \sigma_n\Delta dz_{n,t}$$

We can derive a yield curve iteratively by adding $E[df_{n,t}\,|\,dz_{\nu,t}]$ to $f_{n,t}$ because $\sigma_n\Delta dz_{n,t}$ is a cross-sectional measurement error. This is in line with the usual affine model approach in which each yield equation with measurement errors is specified as:

$$\gamma_t^{(\tau)} = A(\tau) + B(\tau)^T x_t + \epsilon_t^{(\tau)}$$

Matching the volatility terms produces:

$$\begin{aligned}
(df_{n,t} \cdot \Delta)^2 &= \left(-\lambda_t' dz_{\nu,t} + \sigma_{f,n,t}^2 \left(\Delta/2 + \lambda_t' \cdot \vec{\beta_{n,t}}\right)\Delta dt + \sigma_n \Delta dz_{n,t}\right)^2 \\
\iff \sigma_{f,n,t}^2 dt \cdot \Delta^2 &= \lambda_t' \lambda_t dt + (\sigma_n \Delta)^2 dt - 2\sigma_n \Delta \lambda_t' dt \\
\iff \sigma_{f,n,t}^2 \Delta^2 &= \lambda_t' \lambda_t + \sigma_n^2 \Delta^2 \\
\iff \sigma_{f,n,t}^2 &= (\lambda_0 + \lambda_1 \nu_t)'(\lambda_0 + \lambda_1 \nu_t) + \sigma_n^2
\end{aligned}$$

Then, our empirical design becomes:

$$df_{n,t} = -(\lambda_0 + \lambda_1 \nu_t)' dz_{\nu,t} + \kappa_{n,t} \Delta dt + \sigma_n dz_{n,t},$$

where

$$\begin{aligned}
\kappa_{n,t} &= \sigma_{f,n,t}^2 \left(1/2 + (\lambda_0 + \lambda_1 \nu_t)' \cdot \vec{\beta_{n,t}}\right) \\
\sigma_{f,n,t}^2 &= (\lambda_0 + \lambda_1 \nu_t)'(\lambda_0 + \lambda_1 \nu_t) + \sigma_n^2.
\end{aligned}$$

☐ Empirical design: Regression approach

First, define ν_t process on normalization, detrending, and dimension reduction. Second, run the Fama-MacBeth regression to estimate $\{\lambda_0, \lambda_1, \sigma_n\}$based on the assumption that λ_1 a diagonal matrix. Third, regress $dz_{\nu,t}$ on $df_{n,t}$ to obtain $\vec{\beta_{n,t}}$. Finally, estimate standard errors using bootstrapping. If the function form of the yield curve (*f*) is known, we can generate as much cross-sectional data as possible by making Δ infinitely small. It is possible when a yield curve is defined at an infinite number of points (e.g., a real line). Then, we can estimate a large dimension of (λ_0, λ_1). When Δ becomes too small and negligible, we can ignore the term, $\kappa_{n,t}\Delta dt$.

Appendix II. Five-step Procedure to Balance the Panel of Macroeconomic Variables

Step 1. Use the *tcode* from McCracken and Ng (2015) to transform the data.

Step 2. Normalize the outcome from the first step because "observations that are missing are initialized to the unconditional mean based on the non-missing values (which is zero since the data are demeaned and standardized) so that the panel is re-balanced (McCracken and NG, 2015)."

Step 3. Use the generated panel data to obtain factors and loadings before rewriting the missing values with estimates of lambda times factor.

Step 4. Use the standard deviation and mean estimates obtained in the process of normalization in Step 2 to inverse the normalization to revert to the original data form.

Step 5. Repeat Step 2 to step 4 until missing values do not change.

Appendix III. List of macroeconomic variables.

The table lists all 124 macroeconomic variables along with the variable names, descriptions and *tcodes*, following Ludvigson and Ng (2009). The *tcode* column denotes the following data transformation for a series x:

(1) No transformation
(2) Δx_t
(3) $\Delta^2 x_t$
(4) $\log(x_t)$
(5) $\Delta \log(x_t)$
(6) $\Delta^2 \log(x_t)$
(7) $\Delta(x_t/x_{t-1} - 1.0)$

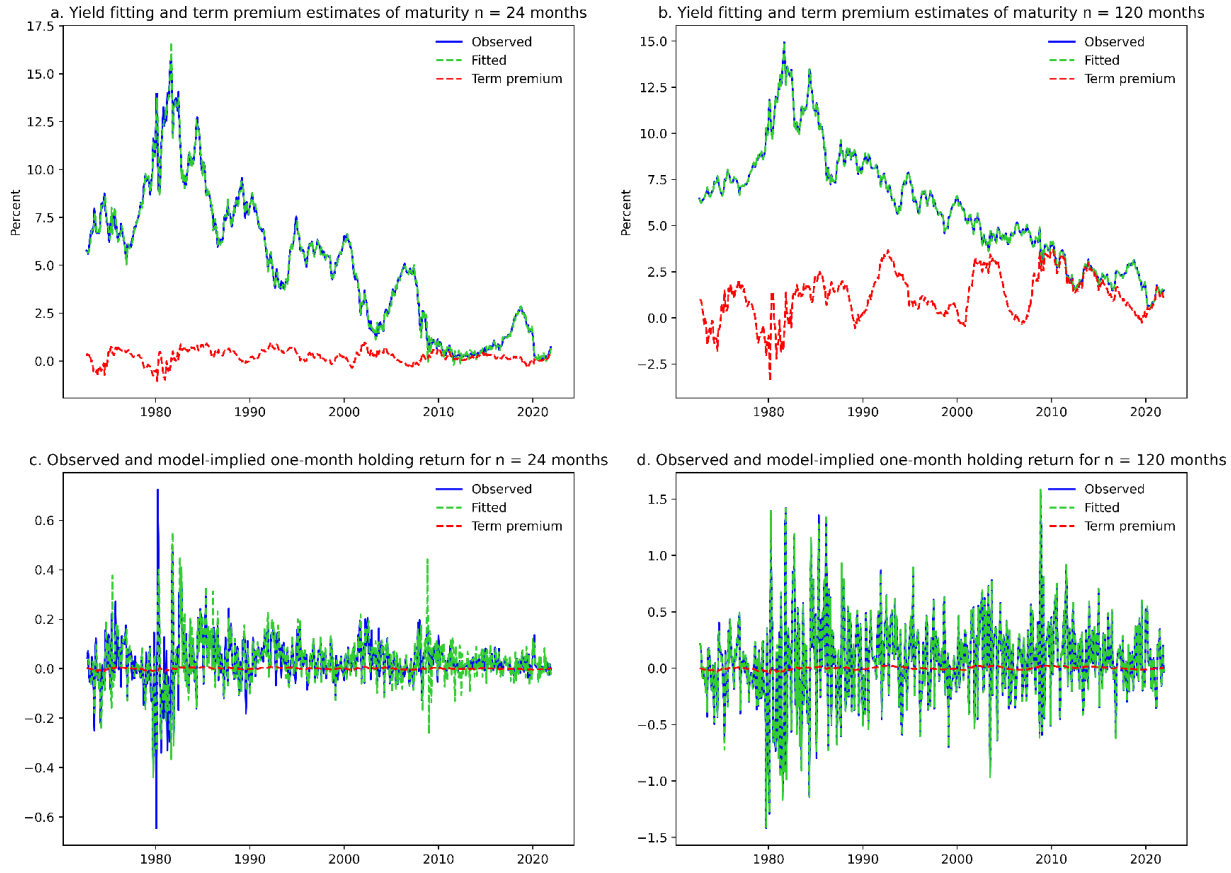
Category	Variable	Description	tcode
Output and Income	IPDCONGD	IP: Durable Consumer Goods	5
	IPFUELS	IP: Fuels	5
	IPBUSEQ	IP: Business Equipment	5
	IPDMAT	IP: Durable Materials	5
	IPNCONGD	IP: Nondurable Consumer Goods	5
	IPFPNSS	IP: Final Products and Nonindustrial Supplies	5
	IPNMAT	IP: Nondurable Materials	5
	IPCONGD	IP: Consumer Goods	5
	IPMAT	IP: Materials	5
	IPFINAL	IP: Final Products (Market Group)	5
	INDPRO	IP Index	5
	RPI	Real Personal Income	5
	IPB51222S	IP: Residential Utilities	5
	IPMANSICS	IP: Manufacturing (SIC)	5
	W875RX1	Real personal income ex transfer receipts	5
	CUMFNS	Capacity Utilization: Manufacturing	2
Labor Market	UNRATE	Civilian Unemployment Rate	2
	DMANEMP	All Employees: Durable goods	5
	USCONS	All Employees: Construction	5
	AWHMAN	Avg Weekly Hours : Manufacturing	1
	UEMP5TO14	Civilians Unemployed for 5-14 Weeks	5
	USTPU	All Employees: Trade, Transportation & Utilities	5
	PAYEMS	All Employees: Total nonfarm	5
	HWIURATIO	Ratio of Help Wanted/No. Unemployed	2
	CES3000000008	Avg Hourly Earnings : Manufacturing	6
	CES2000000008	Avg Hourly Earnings : Construction	6
	CLF16OV	Civilian Labor Force	5
	NDMANEMP	All Employees: Nondurable goods	5
	CES0600000007	Avg Weekly Hours : Goods-Producing	1
	CE16OV	Civilian Employment	5
	SRVPRD	All Employees: Service-Providing Industries	5
	UEMP27OV	Civilians Unemployed for 27 Weeks and Over	5
	UEMPMEAN	Average Duration of Unemployment (Weeks)	2
	MANEMP	All Employees: Manufacturing	5
	UEMPLT5	Civilians Unemployed - Less Than 5 Weeks	5
	CLAIMSx	Initial Claims	5
	UEMP15T26	Civilians Unemployed for 15-26 Weeks	5
	UEMP15OV	Civilians Unemployed - 15 Weeks & Over	5
	USFIRE	All Employees: Financial Activities	5
	USGOOD	All Employees: Goods-Producing Industries	5
	USGOVT	All Employees: Government	5
	USTRADE	All Employees: Retail Trade	5
	CES0600000008	Avg Hourly Earnings : Goods-Producing	6
	USWTRADE	All Employees: Wholesale Trade	5
	AWOTMAN	Avg Weekly Overtime Hours : Manufacturing	2
	CES1021000001	All Employees: Mining and Logging: Mining	5
	HWI	Help-Wanted Index for United States	2
Consumption and Orders	HOUSTMW	Housing Starts, Midwest	4
	HOUSTNE	Housing Starts, Northeast	4
	PERMITS	New Private Housing Permits, South (SAAR)	4
	PERMITW	New Private Housing Permits, West (SAAR)	4
	HOUST	Housing Starts: Total New Privately Owned	4
	PERMIT	New Private Housing Permits (SAAR)	4
	HOUSTW	Housing Starts, West	4
	PERMITMW	New Private Housing Permits, Midwest (SAAR)	4
	PERMITNE	New Private Housing Permits, Northeast (SAAR)	4
	HOUSTS	Housing Starts, South	4
Orders and Inventories	UMCSENTx	Consumer Sentiment Index	2
	DPCERA3M086SBEA	Real personal consumption expenditures	5
	RETAILx	Retail and Food Services Sales	5
	AMDMUOx	Unfilled Orders for Durable Goods	5
	BUSINVx	Total Business Inventories	5
	ISRATIOx	Total Business: Inventories to Sales Ratio	2
	ANDENOx	New Orders for Nondefense Capital Goods	5
	ACOGNO	New Orders for Consumer Goods	5
	CMRMTSPLx	Real Manu. and Trade Industries Sales	5
	AMDMNOx	New Orders for Durable Goods	5

Money and Credit	M1SL	M1 Money Stock	6
	DTCTHFNM	Total Consumer Loans and Leases Outstanding	6
	M2REAL	Real M2 Money Stock	5
	INVEST	Securities in Bank Credit at All Commercial Banks	6
	REALLN	Real Estate Loans at All Commercial Banks	6
	M2SL	M2 Money Stock	6
	NONBORRES	Reserves Of Depository Institutions	7
	TOTRESNS	Total Reserves of Depository Institutions	6
	DTCOLNVHFNM	Consumer Motor Vehicle Loans Outstanding	6
	BUSLOANS	Commercial and Industrial Loans	6
	NONREVSL	Total Nonrevolving Credit	6
	CONSPI	Nonrevolving consumer credit to Personal Income	2
Interest rate and Exchange Rates	AAA	Moody's Seasoned Aaa Corporate Bond Yield	2
	EXJPUSx	Japan / U.S. Foreign Exchange Rate	5
	T10YFFM	10-Year Treasury C Minus FEDFUNDS	1
	TB3MS	3-Month Treasury Bill	2
	GS1	1-Year Treasury Rate	2
	BAAFFM	Moody's Baa Corporate Bond Minus FEDFUNDS	1
	EXCAUSx	Canada / U.S. Foreign Exchange Rate	5
	BAA	Moody's Seasoned Baa Corporate Bond Yield	2
	EXSZUSx	Switzerland / U.S. Foreign Exchange Rate	5
	COMPAPFFx	3-Month Commercial Paper Minus FEDFUNDS	1
	CP3Mx	3-Month AA Financial Commercial Paper Rate	2
	GS5	5-Year Treasury Rate	2
	T1YFFM	1-Year Treasury C Minus FEDFUNDS	1
	TB6SMFFM	6-Month Treasury C Minus FEDFUNDS	1
	FEDFUNDS	Effective Federal Funds Rate	2
	TB3SMFFM	3-Month Treasury C Minus FEDFUNDS	1
	GS10	10-Year Treasury Rate	2
	AAAFFM	Moody's Aaa Corporate Bond Minus FEDFUNDS	1
	TB6MS	6-Month Treasury Bill	2
	EXUSUKx	U.S. / U.K. Foreign Exchange Rate	5
	T5YFFM	5-Year Treasury C Minus FEDFUNDS	1
Prices	CUSR0000SA0L2	CPI : All items less shelter	6
	DDURRG3M086SBEA	Personal Cons. Exp: Durable goods	6
	CPIMEDSL	CPI : Medical Care	6
	WPSFD49207	Producer Price Index by Commodity: Final Demand: Finished Goods	6
	WPSID62	Producer Price Index by Commodity: Intermediate Demand by Commodity Type: Unprocessed Goods for Intermediate Demand	6
	CPIAUCSL	CPI : All Items	6
	CPIAPPSL	CPI : Apparel	6
	DNDGRG3M086SBEA	Personal Cons. Exp: Nondurable goods	6
	CUSR0000SA0L5	CPI : All items less medical care	6
	WPSID61	Producer Price Index by Commodity: Intermediate Demand by Commodity Type: Processed Goods for Intermediate Demand	6
	CUSR0000SAS	CPI : Services	6
	OILPRICEx	Crude Oil, spliced WTI and Cushing	6
	CUSR0000SAD	CPI : Durables	6
	CPITRNSL	CPI : Transportation	6
	PCEPI	Personal Cons. Expend: Chain Index	6
	DSERRG3M086SBEA	Personal Cons. Exp: Services	6
	CPIULFSL	CPI : All Items Less Food	6
	WPSFD49502	Producer Price Index by Commodity: Final Demand: Personal Consumption Goods	6
	PPICMM	PPI: Metals and metal products	6
	CUSR0000SAC	CPI : Commodities	6
Stock Market	S&P: indust	S&P's Common Stock Price Index: Industrials	5
	S&P div yield	S&P's Composite Common Stock: Dividend Yield	2
	VIXCLSx	CBOE S&P 100 Volatility Index: VXO	1
	S&P 500	S&P's Common Stock Price Index: Composite	5
	S&P PE ratio	S&P's Composite Common Stock: Price-Earnings Ratio	5
Others	TWEXAFEGSMTHx	Nominal Major Currencies U.S. Dollar Index (Goods Only)	5
	BOGMBASE	St. Louis Adjusted Monetary Base	6

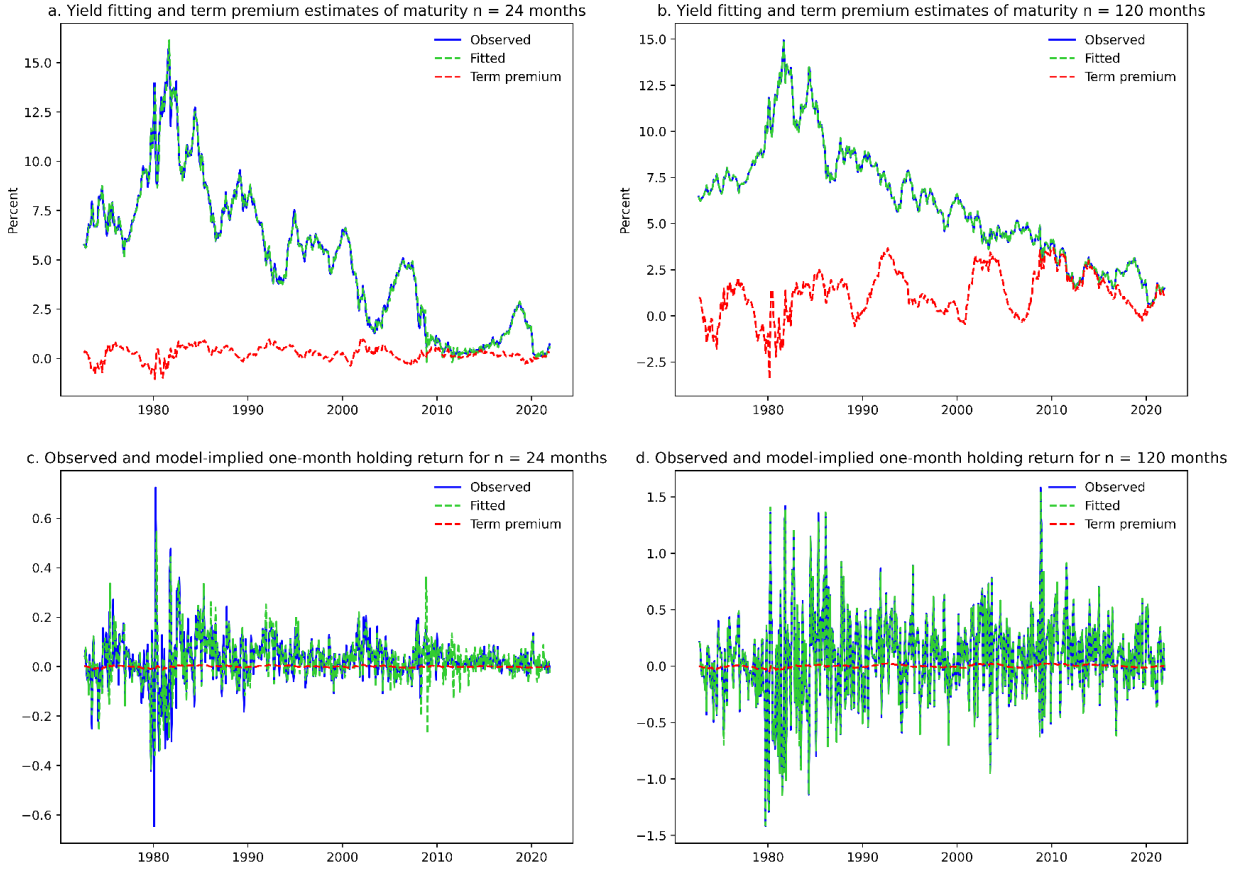
Appendix IV. Time-series fitting and term premia of model-implied yields using four and five yield factors.

This figure plots the yield fitting and term premium estimates, as well as predictability for one-month holding period excess returns of zero-coupon yield curve data for Treasuries with two- and ten-year maturities, as observed and implied by the ACM model and our proposed model using four yield factors (Panel A) and five yield factors (Panel B). The sample data are constructed using Liu and Wu's (2021) yield curve data set. The sample period is from 1972:8 to 2021:12. For both panels, solid lines represent observed yields and returns, dashed green lines represent model-implied yields and returns, and dashed red lines represent the model-implied term premia.

Panel A. Time-series fit and term premia of model-implied yields using four yield factors.

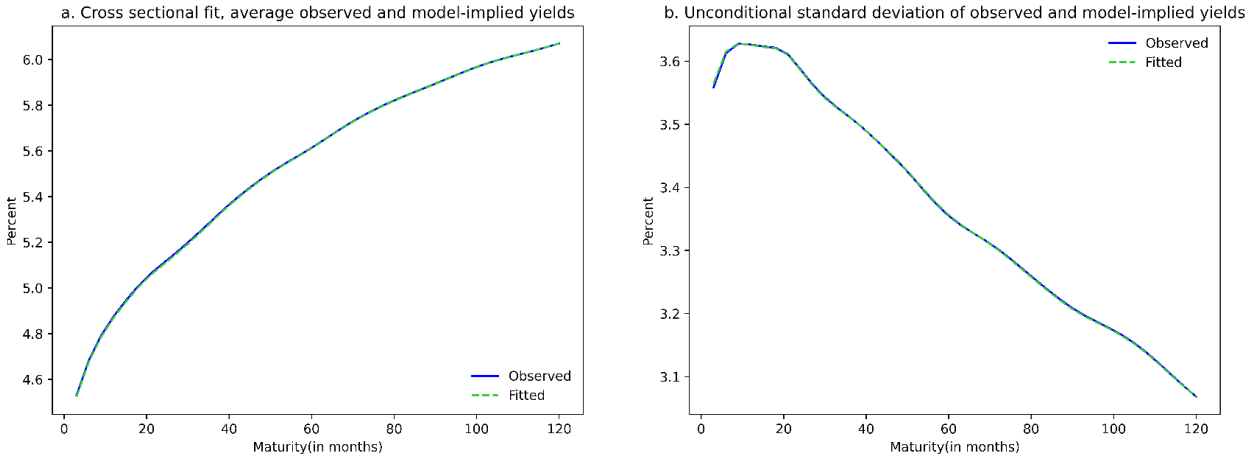


Panel B. Time-series fit and term premia of model-implied yields using five yield factors.



Appendix V. Cross-sectional fit of model-implied yields using four and five yield factors.
The figures plot the cross-sectional fit of the yields generated by using our proposed model with four (Panel A) and five factors (Panel B). The sample data are constructed using Liu and Wu's (2021) yield curve data set. The sample period is from 1972:8 to 2021:12. For both panels, the graph on the left-hand side plots the unconditional means while the graph on the right-hand side plots the unconditional standard deviations of the observed and model-implied yields.

Panel A. Cross-sectional fit of model-implied yields using four yield factors.



Panel B. Cross-sectional fit of model-implied yields using five yield factors.