# Forecasting with the term structure: The role of no-arbitrage restrictions

*Gregory R. Duffee**

Johns Hopkins University

First draft: October 2007
This Draft: January 25, 2011

## ABSTRACT

No-arbitrage term structure models impose cross-sectional restrictions among yields and can be used to impose dynamic restrictions on risk compensation. This paper evaluates the importance of these restrictions when using the term structure to forecast future bond yields. It concludes that no cross-sectional restrictions are helpful, because cross-sectional properties of yields are easy to infer with high precision. Dynamic restrictions are useful, but can be imposed without relying on the no-arbitrage structure. In practice, the most important dynamic restriction is that the first principal component of Treasury yields follows a random walk. A simple model built around this assumption produces out-of-sample forecasts that are more accurate than those of a variety of alternative dynamic models.

*Voice 410-516-8828, email duffee@jhu.edu. Address correspondence to 440 Mergenthaler Hall, 3400 N. Charles St., Baltimore, MD 21218. Thanks to Philippe Mueller and seminar participants at NYU, Penn State, Ohio State, Johns Hopkins, the Spring 2009 Adam Smith Asset Pricing Conference, the Federal Reserve Bank of New York, and Wharton.

# 1 Introduction

No-arbitrage affine term structure models are rapidly becoming important forecasting tools. In particular, Gaussian versions are employed by Duffee (2002), Dai and Singleton (2002), and Christensen, Diebold, and Rudebusch (2010) to predict Treasury yields, by Cochrane and Piazzesi (2008) to predict excess bond returns, and by Ang and Piazzesi (2003) to predict macroeconomic activity. This literature argues that models satisfying no-arbitrage should produce more accurate forecasts than models that do not impose such restrictions.

Affine term structure models are linear factor models of the term structure. Like other standard factor models, they have a time series and a cross-sectional component. The former is a description of the dynamics of a low-dimensional vector of factors. The latter is a linear mapping from factors to the yield on an $m$-maturity bond. No-arbitrage implies the existence of an equivalent-martingale measure, which imposes the restrictions of Duffie and Kan (1996) on this cross-sectional mapping.

Thus if our goal is to forecast anything other than the factors themselves, the value of Duffie-Kan restrictions appears obvious: a relatively small number of parameters determines the entire time-$t$ shape of the yield curve conditional on the time-$t$ factors. If we do not impose no-arbitrage on the factor model, the mapping from factors to yields is much more flexible. Put differently, no-arbitrage reduces substantially the number of cross-sectional free parameters. Standard economic intuition tells us that out-of-sample forecasting accuracy should improve as long as the Duffie-Kan restrictions are satisfied in the data.

This logic is incorrect. The first major contribution of this paper is to show that Duffie-Kan restrictions are unnecessary to estimate the cross-sectional mapping. The intuition is straightforward. If we take literally the assumption that yields are linear functions of some factors (and a constant), those factors are also linear combinations of yields. For example, if $n$ factors determine yields, the state vector can be rotated to equal the first $n$ principal components of the term structure. To determine the cross-sectional mapping from the factors to an $m$-maturity yield, simply regress the yield on the factors. There is no estimation error in the coefficients because the regression $R^2$s are one.

In practice, a low-dimensional factor representation of yields does not exactly hold, hence $R^2$s are not quite one. Therefore empirical applications of term structure models add measurement error to yields. But with a reasonable choice of the number of factors (three is sufficient), variances of measurement errors are tiny relative to variances of yields. Typical cross-sectional $R^2$s are around 0.999. Hence cross-sectional loadings of yields on factors are estimated with extremely high precision using ordinary least-squares. The largest plausible deviation between yields fitted to OLS estimates of the loadings and yields fitted to true

loadings is only a few basis points.

Some readers have incorrectly interpreted this argument as meaning "no-arbitrage holds so strongly in the data that it need not be imposed." Instead, a *linear factor model* holds so strongly in the data that no cross-sectional restrictions are necessary to infer cross-sectional behavior. This does not imply that the Duffie-Kan restrictions are irrelevant. If the restrictions are inconsistent with the true cross-sectional patterns in the data, imposing them will produce different, and presumably worse, estimates of the cross section than those produced without imposing the restrictions. This conclusion also applies to models that impose cross-sectional restrictions that are not derived from no-arbitrage, as in the dynamic Nelson-Siegel model of Diebold and Li (2006). Cross-sectional restrictions bite only if they contradict the true linear factor model.

The application of this result to forecasting picks up where Joslin, Singleton, and Zhu (2011) ends. They show that for Gaussian models that impose no-arbitrage but do not otherwise restrict risk premia dynamics, Duffie-Kan restrictions are irrelevant to estimating factor dynamics. The results here show that the restrictions do not help estimate cross-sectional mappings from factors to yields. In combination, these two conclusions imply that the mere existence of an equivalent-martingale measure contributes nothing to Gaussian term structure estimation and forecasting.

However, the Joslin et al. (2011) result does not apply to a rapidly growing literature that uses no-arbitrage as a framework to impose additional restrictions on the factors. In a Gaussian setting, an equivalent-martingale measure is specified and its properties are parametrically linked to the properties of the physical measure. The result is a Gaussian linear factor model of the term structure that has both restrictions on factor dynamics and Duffie-Kan restrictions on the cross section. The former restrictions are equivalent to restrictions on the dynamics of risk premia.

The second major contribution of this paper is to show that empirically valuable restrictions on Gaussian factor dynamics can be imposed without relying on a researcher's ability to intuit the correct functional form of risk compensation. I develop a parsimonious three-factor dynamic term structure model that does not impose an equivalent-martingale measure.

The factors are the first three principal components of yields. The model imposes a random walk on the first principal component, and imposes stationarity on the other two components. It has unrestricted mappings from factors to yields, in contrast to no-arbitrage models as well as the dynamic Nelson-Siegel models of Diebold and Li (2006) and Christensen et al. (2010). The model produces out-of-sample forecasts of yields that dominate those of many alternative dynamic models.

The next section evaluates the role of no-arbitrage in estimating cross-sectional relationships among yields. Section 3 applies the conclusions of Section 2 to forecasting future yields. Dynamic restrictions on term structure models are considered in Section 4. Concluding comments are in the final section.

# 2 No-arbitrage and the cross section of yields

This section demonstrates that Duffie-Kan restrictions are unecessary to estimate the cross-sectional properties of Treasury yields. The first subsection describes, in fairly general terms, the no-arbitrage affine framework and its associated restrictions. The second subsection nests no-arbitrage models in a broader class of linear factor models. The third subsection describes the Gaussian special case of the no-arbitrage and linear factor models, which are used extensively in Section 3. The fourth subsection discusses estimation of no-arbitrage and linear factor models. It explains why the usefulness of cross-sectional restrictions depends on the amount of noise (e.g., measurement error) in observed yields. The final subsection shows that empirically, deviations of Treasury yields from a three-factor linear model are very small. The implied amount of noise in yields is not big enough to make any cross-sectional restrictions useful.

## 2.1 A no-arbitrage affine setting

There are no arbitrage opportunities, hence there is an equivalent-martingale measure. Under this measure, assume that the risk-free short rate is an affine function of a length-$n$ state vector $x_t$,

$$r_t = \delta_{0X} + \delta_{1X}' x_t. \tag{1}$$

Also adopt standard perfect market assumptions. There are no trading costs, no asymmetric information, no agents with market power, and no taxes. Hence we rule out a variety of real-world features of financial markets that will be mentioned in Section 2.2. With perfect markets, a zero-coupon bond's price is its payment at maturity discounted by expected short rates over the life of the bond. The expectations are calculated using equivalent-martingale dynamics.

Finally, assume that the equivalent-martingale dynamics of $x_t$ (including behavior at any boundaries) are in the set of dynamics that, when combined with the short-rate equation (1), produces an affine mapping from the state vector to zero-coupon bond yields. Duffie and Kan (1996), building on the work of Vasicek (1977) and Cox, Ingersoll, and Ross (1985), describe continuous-time models in this class. Discrete-time models include the Gaussian

class first explored by Backus and Zin (1994) and a non-Gaussian class characterized by Le, Singleton, and Dai (2010).

Since $x_t$ is a latent vector, normalizations must be imposed to identify the parameters of the equivalent-martingale dynamics and the short-rate equation. Denote an identified parameter vector as $\Phi_X^q$. The makeup of $\Phi_X^q$ depends on the particular model. An example of the elements of $\Phi_X^q$ is presented in Section 2.3, which lays out the workhorse discrete-time Gaussian model. In the general affine case, the mapping from factors to yields is

$$y_t^{(m)} = \alpha_X^{(m)}(\Phi_X^q) + \beta_X^{(m)}(\Phi_X^q)' x_t, \quad x_t \in \Omega_X(\Phi_X^q), \tag{2}$$

where the yield on an $m$-maturity bond is denoted $y_t^{(m)}$. The functions $\alpha_X^{(m)}$ and $\beta_X^{(m)}$ are model-specific. They are calculated using the differential equations of Duffie and Kan (1996) or their difference equation counterparts. I refer to the general technique as the Duffie-Kan recursions. The state vector can take on any value in $\Omega_X$, which is a subset of $\mathbb{R}^n$. This space is determined by the equivalent-martingale dynamics.

To understand the role of no-arbitrage in affine models, it is helpful to transform the state vector. Define the $d$-vector $Y_t$, $d > n$, as a vector of yields on bonds with constant maturities $\mathcal{M} = \{m_1, \ldots, m_d\}$. Stack each yield's affine mapping (2) to express the yield vector as

$$Y_t = A_X(\mathcal{M}, \Phi_X^q) + B_X(\mathcal{M}, \Phi_X^q) x_t. \tag{3}$$

Let $P$ be an $n \times d$ matrix with rank $n$. Use this matrix to express $n$ linear combinations of yields as a function of the state vector,

$$\mathcal{P}_t \equiv PY_t = PA_X(\mathcal{M}, \Phi_X^q) + PB_X(\mathcal{M}, \Phi_X^q) x_t. \tag{4}$$

To simplify notation, the arguments of the $d$-vector $A_X$ and the $d \times n$ matrix $B_X$ are henceforth suppressed. Outside of knife-edge cases, the matrix $PB_X$ is invertible, so $\mathcal{P}_t$ contains the same information as $x_t$. Substitute $x_t$ out of (3) using (4), expressing yields as affine functions of $\mathcal{P}_t$:

$$Y_t = A_P(\mathcal{M}, \Phi_X^q, P) + B_P(\mathcal{M}, \Phi_X^q, P)\mathcal{P}_t, \qquad \mathcal{P}_t \in \Omega_P(\mathcal{M}, \Phi_X^q, P), \tag{5}$$

$$A_P = \left(I_{d\times d} - B_X \left(PB_X\right)^{-1} P\right) A_X, \quad B_P = B_X \left(PB_X\right)^{-1}. \tag{6}$$

In (5), $\Omega_P$ is the subset of $\mathbb{R}^n$ that is the admissible space for $\mathcal{P}_t$. Note that the vector $A_P$ and the matrix $B_P$ ensure the internal consistency of (5). Mathematically, this means premultiplication of $A_P$ and $B_P$ by $P$ produces a zero vector and an identity matrix respec-

tively.

## 2.2 A general linear factor model

There are no arbitrage opportunities. But the absence of arbitrage alone, without the assumption of perfect markets, does not imply (2). The market for Treasury securities is decidedly imperfect. Both trading costs and institutional features of the Treasury market affect Treasury yields. For example, owners of on-the-run Treasury bonds usually have the ability to borrow at below-market interest rates in the repurchase market. Certain Treasury securities trade at a premium because they are the cheapest to deliver in fulfillment of futures contract obligations. Treasury debt is more liquid than non-Treasury debt, which is one reason why Treasury bonds are perceived to offer a "convenience yield" to investors in addition to the yield calculated from price. In a nutshell, returns calculated from bond yields do not necessarily correspond to returns realized by investors. Evidence suggests that these market imperfections can have significant effects on observed yields.[1]

If we are unwilling to assume perfect markets, we can replace (5) with a standard linear factor framework. Yields are affine functions of the state vector, but no cross-bond restrictions are applied to

$$Y_t = a_P + b_P \mathcal{P}_t, \qquad \mathcal{P}_t \in \mathbb{R}^n. \tag{7}$$

Underlying (7) is the assumption that the effects of market imperfections on yields are affine functions of the state. We weaken this assumption in Section 2.4. The only constraint on the coefficients of (7) is that premultiplication of (7) by $P$ must produce $\mathcal{P}_t$ on the right side. Therefore

$$Pa_P = 0, \quad Pb_P = I. \tag{8}$$

There are $d(n+1)$ elements of $a_P$ and $b_P$ in (7) and $n(n+1)$ constraints on these elements in (8). Therefore there are $(d-n)(n+1)$ free cross-sectional parameters.

The mapping (7) describes the cross section of yields $Y_t$. In this sense, it is a term structure model that nests the no-arbitrage model of Section 2.1, which also describes this cross section. The cost of not using the Duffie-Kan recursions is that we give up the ability to determine prices of other fixed-income instruments. The no-arbitrage model is a model of the cross section of fixed income. It can be used to price all claims contingent on these yields, such as coupon bonds and bond options. Equation (7) only prices bonds in terms

[1]The first academic evidence appears to be Park and Reinganum (1986). Early research focused on prices of securities with remaining maturities of only a few weeks or months. Duffee (1996) contains evidence and references to earlier work. Evidence of market imperfections at longer maturities is in Krishnamurthy (2002), Greenwood and Vayanos (2010), and Krishnamurthy and Vissing-Jorgensen (2007).

of other bonds. However, this limitation is irrelevant from the perspective of much of the empirical dynamic term structure literature because of the literature's exclusive focus on yield dynamics.

## 2.3 A discrete-time Gaussian example

Discrete-time Gaussian models are used extensively in Section 3. Using a version of the identification scheme of Joslin et al. (2011), the state vector's equivalent-martingale dynamics are

$$x_{t+1} = \text{diag}(g)x_t + \Sigma_X \epsilon^q_{t+1}, \quad \epsilon^q_{t+1} \sim MVN(0, I). \tag{9}$$

The notation diag($g$) denotes a diagonal matrix with the vector $g$ along the diagonal. This vector consists of distinct real values, none of which equals one. By construction, they are the eigenvalues of diag($g$). More general specifications of eigenvalues are considered in Joslin et al. (2011). The matrix $\Sigma_X$ is lower triangular. The short-rate equation (1) is specialized to the case where the short rate is the sum of a constant and the elements of the state vector,

$$r_t = \delta_{0X} + \iota' x_t, \tag{10}$$

where $\iota$ is an $n$-vector of ones.

The physical dynamics of $x_t$ are not modeled here because they are irrelevant to pricing. Thus the parameter vector contains only the parameters of (10) and (9),

$$\Phi^q_X = \{\delta_{0X}, g, \text{vech}(\Sigma_X)\}, \tag{11}$$

which has $1+n+n(n+1)/2$ elements. The functional forms of the no-arbitrage yield mapping (2) are in the appendix.

Contrast the two mappings from the state vector $\mathcal{P}_t$ to the yield vector $Y_t$, given by (5) and (7). The linear factor version has $(d-n-1)(n+1) - n(n+1)/2$ additional free parameters. Put differently, the Duffie-Kan recursions of the Gaussian model impose $(d-n-1)(n+1) - n(n+1)/2$ restrictions on the cross section of $d$ yields.

The intuition behind the number of restrictions is straightforward. In any no-arbitrage model with $n$ shocks, the prices of $n$ assets determine the compensation for each of the $n$ shocks, given the magnitude of the shocks and the short rate. The $n(n+1)/2$ parameters of $\Sigma_X$ determine the magnitudes of the $n$ shocks to the term structure. Here we do not observe the short rate (unless it is included in $Y_t$), therefore we must observe the price of an additional bond in order to pin it down. Put differently, if we know the parameters of $\Sigma_X$ and observe prices (yields) of only $n+1$ bonds, no-arbitrage has no bite relative to a

more general linear factor model.[2] There is a one-to-one mapping from the linear factor model parameters to the no-arbitrage model parameters. Each additional bond adds $n+1$ overidentifying restrictions because the additional bond must be priced consistently with the initial $n+1$ bonds.

## 2.4 Some cross-sectional estimation intuition

Consider a empirical setting where we observe a $d$-vector of constant-maturity yields $\widetilde{Y}_t, t = 1, \ldots, T$. We want to estimate parameters of both an $n$-factor no-arbitrage model and an $n$-factor linear factor model. We also want to test the null hypothesis that overidentifying restrictions implied by the Duffie-Kan recursions are consistent with the data. Formally, the null hypothesis is

$$\begin{aligned} H0: \quad & a_P = A_P\left(\mathcal{M}, \Phi_X^q, P\right), \\ & b_P = B_p\left(\mathcal{M}, \Phi_X^q, P\right), \\ & \mathcal{P}_t \in \Omega_P\left(\mathcal{M}, \Phi_X^q, P\right). \end{aligned} \tag{12}$$

The maintained hypothesis is that the linear factor model (7) holds.

Estimation and testing critically depend on the relation between observed yields $\widetilde{Y}_t$ and yields in the model, $Y_t$. A standard setting is where observed yields differ from model-implied yields owing to some idiosyncratic component in yields

$$\widetilde{Y}_t = Y_t + \eta_t, \qquad E(\eta_t) = 0,\ E(x_t \eta_{t+j}') = 0 \ \forall\ j. \tag{13}$$

The observable proxy for the state vector is

$$\widetilde{\mathcal{P}}_t \equiv P\widetilde{Y}_t = \mathcal{P}_t + P\eta_t. \tag{14}$$

The stochastic component $\eta_t$ is treated as measurement error when the no-arbitrage framework of Section 2.1 is used. With the more general linear factor framework, $\eta_t$ may be interpreted broadly as idiosyncratic components of yields. These may be the product of preferred habitats, special repurchase rates, or variations in liquidity. The additional structure placed on this noise determines how the models should be estimated.

To develop intuition for estimation and testing, first consider the case where $\eta_t$ is identically zero, or

[2]This is a slight overgeneralization. No-arbitrage imposes some inequality constraints. For example, in a one-factor Gaussian model with two observed bond yields, no-arbitrage requires that the relation between the factor and the yields is the same (positive or negative) for both bonds.

**Noise assumption 1**

$$\widetilde{Y}_t = Y_t \; \forall \; t. \tag{15}$$

There is no measurement error or other idiosyncratic components to yields. With this assumption, cross-sectional estimation and testing are trivial, as shown by the following two propositions.

**Proposition 1.** If the linear factor model cross-sectional mapping (7) is correct and observed yields satisfy (15), exactly one parameter vector satisfies the mapping.

Proof. With (15), the state vector $\mathcal{P}_t$ is observable. The proposition stipulates that (7) holds. Replacing $Y_t$ on the left of (7) with $\widetilde{Y}_t$ produces an affine equation in terms of observables. The parameters of the equation are estimated without error by OLS regressions since the $R^2$s of the regressions are all one. Denote these estimates by $\hat{a}_P^{OLS}$ and $\hat{b}_P^{OLS}$. Any alternative choice of $a_P$ or $b_P$ violates (7) for some pair $\{\widetilde{Y}_t, \mathcal{P}_t\}$ in $t = 1, \ldots, T$.[3]

If these OLS regressions do not have $R^2$s of one, then either the general linear factor model or the observation assumption (15) is false. The next proposition states the same conclusion for the no-arbitrage model.

**Proposition 2.** If the cross-sectional mapping of the no-arbitrage model is correct and observed yields satisfy (15), exactly one parameter vector satisfies the mapping.

Proof. Follows Proposition 1. The parameter vector is calculated by replacing $Y_t$ with $\widetilde{Y}_t$ in (5) and numerically solving for $\Phi_X^q$ by minimizing squared errors. Since the parameter vector is identified by definition, errors are identically zero for a single value of the vector. Denote this vector by $\hat{\Phi}_X^{q,ML}$, indicating a maximum likelihood (ML) solution. Conditional on the realized state vectors, the likelihood of observing the time series of $\widetilde{Y}_t$ is one when $\Phi_X^q = \hat{\Phi}_X^{q,ML}$, and equals zero for any alternative candidate parameter vector. Once the parameter vector is determined, the admissible space $\Omega_P$ of $\mathcal{P}_t$ can be calculated.

If there is no parameter vector for which errors are identically zero, or if the time series $\mathcal{P}_t, t = 1, \ldots, T$ is not in $\Omega_P$, then either the no-arbitrage model or the observation assumption (15) is false.

The combination of these propositions is

[3]Assume that the length of the time series $T$ is sufficient to identify each set of parameters. For example, $T = 1$ does not allow us to identify separately the constant term $a_P$ and the factor loadings $b_P$.

**Corollary 1.** If the cross-sectional mapping of the no-arbitrage model is correct and observed yields satisfy (15), then the null hypothesis (12) is identically satisfied by

$$\begin{aligned} H0: \quad \hat{a}_P^{OLS} &= A_P\left(\mathcal{M}, \hat{\Phi}_X^{q,ML}, P\right), \\ \hat{b}_P^{OLS} &= B_p\left(\mathcal{M}, \hat{\Phi}_X^{q,ML}, P\right). \end{aligned} \tag{16}$$

The corollary tells us that if the Duffie-Kan overidentifying restrictions are exactly satisfied in the data, imposing them does not improve the estimation accuracy of the cross-sectional mapping from factors to yields. This is not a statement that the overidentifying restrictions are weak in some way. Instead, the point is that with a linear factor model, equivalent-martingale dynamics are unnecessary to estimate cross-sectional mappings.

This strict conclusion relies on the lack of noise in observed yields. With sufficient noise in observed yields, Duffie-Kan restrictions presumably would be helpful in estimating cross-sectional mappings. However, the following empirical evidence shows that there simply isn't enough noise to appreciably alter this conclusion.

## 2.5 The cross-sectional behavior of Treasury yields

In practice, how closely do Treasury yields adhere to an exact linear factor model? The answer to this question presumably depends on the data sample. Empirical research typically uses month-end observations of zero-coupon bond yields that are interpolated from coupon bond yields. I therefore focus on these data.

I use yields on eight zero-coupon bonds. The maturities are three months, one through five years, ten years, and fifteen years. The three-month yield is the T-bill yield. Yields from one through five years are constructed by Center for Research in Security Prices (CRSP). Yields on the ten and fifteen years bonds are constructed by staff at the Federal Reserve Board. The sample is 456 month-end observations from January 1972 through December 2009. Earlier data are available, but long-term bonds were not a prominent component of the Treasury market until late 1971.

Stack the eight observed yields at month-end $t$ in $\widetilde{Y}_t$. Following Litterman and Scheinkman (1991), I examine how well the cross section of yields is explained by three factors. The matrix $P$ is chosen to correspond to the loadings of first three principal components of $\widetilde{Y}_t$. The empirical counterpart to (7) is

$$\widetilde{Y}_t = a + b\widetilde{\mathcal{P}}_t + \hat{\eta}_t. \tag{17}$$

I estimate the eight equations separately with OLS. One of the advantages of OLS estimation

is that the internal consistency requirements of (8) are automatically satisfied. Results are in Table 1.

The main message of this table is that the three-factor linear model is an almost exact fit to these data. The mean $R^2$ across the eight regressions is 0.9993. Residual standard deviations range from five to twelve basis points of annualized yields. First-order serial correlations of residuals are high. By lag 19, the serial correlation has effectively disappeared. Therefore the standard errors are adjusted for generalized heteroskedasticity and 18 lags of moving-average residuals.

Interpolation of zero-coupon yields from coupon bond yields introduces measurement error in yields. Bekaert, Hodrick, and Marshall (1997) estimate that, for maturities between one and five years, standard deviations of such measurement error are in the range of six to nine basis points. It is worth noting that the standard deviations of residuals for the artificially-constructed bonds in Table 1 are in the neighborhood of these standard deviations. However, since linear combinations of any measurement error also appear on the right side of (17), we cannot jump to the conclusion that this three-factor model holds exactly up to measurement error induced by interpolation.

Because the residuals are so small, the coefficients of (7) are estimated with extremely high precision. To put this precision into an economically meaningful context, I look at worst-case scenarios for differences between fitted yields using OLS estimates of (17) and yields implied by the unknown true coefficients of (17). For these worst-case scenarios, assume that the true regression coefficients for bond $i$ are all two standard errors away from the point estimates reported in Table 1. For example, the true constant term for the three-year bond is either $-7.8$ basis points or 0.8 basis points instead of the estimated 3.5 basis points. The worst-case scenarios all have realizations of the factors $\widetilde{\mathcal{P}}_t$ that are two sample standard deviations away from their means. For bond $i$, the worst-case scenario produces an absolute difference between OLS-fitted and "true" yields of

$$|\text{worst case}|_i = 2\left(\text{SE}_i(\text{constant term}) + 2\sum_{i=j}^{3}\text{SE}_i(\text{coef on } PC_i)\ \text{Std}(\mathcal{P}_{i,t})\right). \tag{18}$$

Across the eight bonds, the worst of the worst-case scenarios produces a maximum absolute difference between OLS-fitted and "true" yields of only seven basis points. Almost all of this uncertainty is created by sampling error in the regressions' constant terms.

The strikingly high $R^2$s in Table 1 are at the heart of this conclusion. To ensure that $R^2$s close to one are the norm, not the exception, I estimate the same regressions over subsamples. There are 337 overlapping samples of 120 months in the 1972 through 2009 period. For each of these ten-year periods, I construct principal components of the eight

bond yields, then regress individual yields on thse first three components. This exercise produces, for each bond, 337 $R^2$s, residual standard deviations, and residual first-order serial correlations. Table 2 reports the means of these statistics. It also reports the minimum $R^2$s and the maximum residual standard deviations and residual serial correlations. Because the numbers do not differ much from one bond to another, results are reported for only five of the eight simulated bond yields.

The clear conclusion to draw from Table 2 is that the three-factor linear model fits each of these subsamples almost perfectly. The mean $R^2$s are in the range of 0.997 to 0.999. The lowest $R^2$ across all subsamples and all bonds is 0.994. Residual standard deviations are in the range of five to ten basis points; the highest across all subsamples and bonds is twelve basis points. The serial correlations are around 0.5 to 0.7, although there are subsamples with particularly high serial correlations for the longest-maturity bonds (in excess of 0.8).

Given these results, it is hard to imagine that no-arbitrage models are much of a help in fitting the cross-section of yields. Monte Carlo simulations in Section 3.3 confirm the cross-sectional irrelevance of the Duffie-Kan restrictions, but the simulation evidence does nothing more than verify an obvious conclusion. It is clear from the extreme case of mismeasurement in the full sample, as calculated in (18). Imagine that somehow we have discovered the true functional form of the no-arbitrage model that describes yields over this sample, and our estimates of this model's parameters are infinitely precise. Using this no-arbitrage model, the best improvement in fit relative to cross-sectional regressions is only seven basis points, an amount that can be lost in the measurement error of zero-coupon bond yields.

# 3 Forecasting

The Duffie-Kan restrictions, and no-arbitrage restrictions in general, are inherently cross-sectional. Nonetheless, Duffie-Kan restrictions may be useful in estimating physical dynamics, even when they are unneeded to estimate cross-sectional relationships. The reason is that physical dynamics share features of equivalent-martingale dynamics. Thus equivalent-martingale parameters estimated from the cross section can help pin down physical dynamics. The closer the links between these two sets of dynamics, the more informative the cross section is about the time series.

But as with cross-sectional estimation, whether the restrictions are useful in practice is ultimately an empirical question. This section asks whether the Duffie-Kan restrictions on Gaussian models improve out-of-sample forecasts of bond yields. Any dynamic term structure model can be used for forecasting future yields. However, the literature focuses almost exclusively on Gaussian models because of their flexibility in fitting observed variations in

expected excess bond returns.

This section shows that if the true data-generating process is Gaussian and satisfies the Duffie-Kan restrictions, imposing the restrictions has no effect on forecast accuracy. According to Monte Carlo simulations, out-of-sample forecasts produced by estimating a simple linear factor Gaussian model are very close to forecasts produced by estimating an essentially affine Gaussian model. Differences in root mean squared errors of these forecasts are measured in hundredths of basis points.

The first subsection describes these two dynamic term structure models. The second explains how the models are estimated and distinguishes between the contributions of Joslin et al. (2011) and the current paper. The third subsection presents the Monte Carlo evidence. The fourth asks a related question: is the cross section of Treasury yields consistent with a three-factor essentially affine model? It concludes that deviations from this model are economically very small, but statistically strong.

## 3.1 Two Gaussian dynamic models

As in the analysis of the cross section, treat the $n$ linear combinations of yields $\mathcal{P}_t$ as the state vector. Since this is the same vector that determines the cross section of yields, hidden factors that play a role in the models of Duffee (2011) and Joslin, Priebsch, and Singleton (2010) are ruled out.

We work with two Gaussian models. One is an unrestricted essentially affine model, which imposes Duffie-Kan restrictions. The other is a linear factor model, which ignores no-arbitrage and nests the former model. The models have the same physical dynamics of the state vector, given by

$$\mathcal{P}_{t+1} = \mu_P + K_P \mathcal{P}_t + \Sigma_P \epsilon_{t+1}. \tag{19}$$

The matrix $\Sigma_P$ is lower triangular. The linear factor model consists of (19) and the cross-sectional mapping (7). This combination is a complete dynamic description of the evolution of the constant-maturity yield vector $Y_t$.

No-arbitrage connects physical dynamics to equivalent-martingale dynamics. The unrestricted essentially affine version Gaussian model, introduced in Duffee (2002), has the weakest connection. The two measures share only the conditional volatility of shocks. Therefore $\mu_P$ and $K_P$ are free parameters. Restricted essentially affine models impose links between conditional expectations under the physical and equivalent-martingale measures, producing restrictions on $\mu_P$ and $K_P$. Restricted models are considered in Section 4.

Following the parameter identification approach in Joslin et al. (2011), the volatility

matrix $\Sigma_P$ is

$$\Sigma_P = \left(PB_X\Sigma_X\Sigma_X'B_X'P'\right)^{1/2}, \tag{20}$$

where $\Sigma_X$ is the volatility matrix of the equivalent-martingale dynamics (9). Therefore the free parameters of the unrestricted essentially affine model are

$$\Phi_P = \{\Phi_X^q, \mu_P, \text{vec}(K_P); P, \mathcal{M}\}. \tag{21}$$

The parameter vector $\Phi_X^q$ is defined by (11). The parameters $\mu_P$ and $K_P$ depend on the factor rotation, which in turn depends on the exogenously-specified matrix $P$ and maturities $\mathcal{M}$.

## 3.2 Forecasting with the two models

Return to the empirical setting of Section 2.4. We want to forecast, as of the end of the sample $T$, the bond yield vector as of time $T+s$. The forecasting tools are the two dynamic Gaussian models of Section 3.1. The structure placed on the noise $\eta_t$ in (13) determines how we should estimate parameters and forecast yields. We follow Joslin et al. (2011) by adopting the ad-hoc, but particularly convenient assumption

**Noise assumption 2**

$$P\eta_t = 0 \;\forall\; t, \quad L\eta_t \sim N(0, \sigma_\eta^2 I), \quad E\left(\eta_t\eta_{t+s}'\right) = 0 \;\forall\; s \neq 0, \tag{22}$$

where $L$ is a $(d-n)\times d$ matrix with rank $d-n$ that is linearly independent of the matrix $P$. This assumption implies that the state vector $\mathcal{P}_t$ is observed without noise by the econometrician. The assumption that $\eta_t$ is independent across time is at odds with the serial correlations reported in Tables 1 and 2, but for our purposes little is gained with a more complicated error structure.

Begin with estimation of the unrestricted essentially affine model. The major result of Joslin et al. (2011) is that in this setting, conditional ML estimates of $\mu_P$ and $K_P$ in (19) correspond to their unrestricted vector-autoregression (VAR) estimates. Therefore the Duffie-Kan restrictions are irrelevant to forecasting the state vector.

Their result may appear obvious from the representation of the free parameter vector (21) because $\mu_P$ and $K_P$ do not appear in the equivalent-martingale dynamics. However, although this parameter separation is necessary for their result, it is not sufficient. There are two other important ingredients. The first is observability of the state vector. When it is not observable, maximum likelihood uses filtering to infer the state's physical dynamics. The

information in a given bond's yield about these physical dynamics depends on the amount of noise in the yield, which is estimated using the cross-sectional mapping from the state vector to yields. The second is Gaussian physical dynamics. If volatilities were state-dependent, the intuition of weighted least squares would apply, preventing separation of the conditional mean and conditional variance parameters in ML estimation.

Given the VAR estimates of $\mu_P$ and $K_P$, denote the time-$t$ forecasts of the time-$(t+s)$ state vector as $E^{VAR}(\mathcal{P}_{t+s}|\mathcal{P}_t)$. The unrestricted essentially affine forecast of time-$(t+s)$ yield vector is then

$$E\left(Y_{t+s}|\mathcal{P}_t\right) = A_P(\mathcal{M}, \hat{\Phi}_X^{q,ML}, P) + B_P(\mathcal{M}, \hat{\Phi}_X^{q,ML}, P)E^{VAR}\left(\mathcal{P}_{t+s}|\mathcal{P}_t\right), \tag{23}$$

where ML estimation of the parameter vector $\Phi_X^q$ follows Joslin et al. (2011).

Estimation of the linear factor model is simpler. When noise in yields is characterized by (22), conditional ML estimation of the physical dynamics (19) is equivalent to VAR estimation. Similarly, ML estimation of the cross-sectional parameters $a_P$ and $b_P$ in (7) is equivalent to OLS regressions of observed yields on the observed state vector. As noted in Section 2.5, OLS estimation automatically enforces the internal consistency constraints of (8). Forecasts produced with the linear factor model have the same form as (23). The only difference is that $A_P$ and $B_P$ are replaced with estimates of $a_P$ and $b_P$.

Since $\mathcal{P}_t$ is a linear combination of yields, it is not obvious from (23) how important the estimates of $A_P$ and $B_P$ (or, in the linear factor model case, $a_P$ and $b_P$) are to forecasting. After all, if we are only interested in forecasting future values of the state vector itself, these cross-sectional estimates play no role. For example, if our goal is to forecast the first $n$ principal components of yields, we can simply estimate an unrestricted VAR of these components. This is the focus of Joslin et al. (2011).

However, practical forecasting problems often involve predicting particular yields or yield spreads. For these problems, knowledge of the cross-sectional mapping from $\mathcal{P}_t$ to $Y_t$ is critical. An example of forecasting with a three-factor unrestricted essentially affine model illustrates the importance of the mapping. Using full sample of Treasury yields described in Section 2.5, I estimate a VAR for the first three principal components. The parameters of the equivalent-martingale measure are estimated with conditional ML. I then pick a particular value of $\mathcal{P}_t$ and generate a twelve-month-ahead forecast of the term structure.[4]

The role of the cross-sectional mapping in this forecast is shown in Figure 1. The figure displays three alternative twelve-month-ahead forecasts that differ in the cross-sectional mapping. All three mappings satisfy the Duffie-Kan restrictions, hence they differ in the

[4]Each element of $\mathcal{P}_t$ is set to one-half of its unconditional standard deviation above its mean.

parameters of (11). The solid line uses the actual ML estimates of this mapping, while the other two lines use alternative parameterizations. The differences in forecasts can be substantial. For example, forecasts of the ten-year yield disagree by about one percentage point. Since all of these forecasts are compatible with the VAR model of principal components, the VAR model alone cannot determine the expected future ten-year yield.

Thus estimates of the cross-sectional mapping are crucial for forecasting. Recall the main argument of Section 2, which is that unrestricted estimates of the mapping are so accurate that there is no need to impose any cross-sectional restrictions. The argument carries over to the case of forecasting with the two Gaussian models here, since they differ only in their cross-sectional mappings. The Monte Carlo simulations that follow verify this conclusion.

## 3.3 Monte Carlo analysis

What are the practical effects on estimation and forecasting of imposing the Duffie-Kan restrictions? Monte Carlo simulations help answer this broad question. The simulations shed light on three specific issues. First, returning to the question examined in Section 2, how do the restrictions affect estimates of the cross-sectional mapping from factors to yields? Second, how do the restrictions affect out-of-sample yield forecasts? Third, does the accuracy of these forecasts depend on whether the restrictions are imposed?

The message throughout this paper is that Duffie-Kan restrictions are unneeded. To make this point as convincing as possible, the Monte Carlo simulations are tilted in the direction of finding a role for the restrictions. The simulated data samples are shorter than those typically used in empirical work, which reduces the information that can be gleaned from the data. Similarly, the simulated noise in yields is more volatile and more persistent than we see in Treasury yields. Finally, various "true" data-generating processes are used to see if the results are sensitive to this choice.

### 3.3.1 Comparing cross-sectional fits

The first Monte Carlo exercise looks at estimates of the cross-sectional mapping from factors to yields. The monthly data-generating process is an unrestricted essentially affine three-factor Gaussian model. In order to choose sensible parameters for this process, the model is first estimated with conditional maximum likelihood on the full data sample described in Section 2.5. Estimation follows the procedure of Section 3.2. The parameters of the "true" data-generating process are set equal to the parameter estimates.

The data-generating process specifies that yields are contaminated by measurement error. Measurement error is mean zero, independent across bonds, and serially correlated. The first-

order serial correlation is 0.7 and the unconditional standard deviation is 20 annualized basis points. This is a relatively large amount of noise. Recall from Tables 1 and 2 that deviations from an exact three-factor mdel of Treasury yields are around 10 basis points with a monthly serial correlation of about 0.6.

A simulated data sample is 120 monthly observations of eight bond yields. The maturities are three months, one through five years, ten years, and fifteen years. These are the same maturities of the yields in the actual Treasury data sample. The two Gaussian term structure models of Section 3.1 are estimated on each simulated sample. The parameters of the physical dynamics (19) depend on the matrix $P$ because $P$ determines the state vector $\mathcal{P}_t$. The matrix is set to the loadings of the first three principal components of observed bond yields in the 120-month simulated sample.

Given a data sample, the two Gaussian models are estimated with the procedure of Section 3.2, which relies on the specification of noise (22). The matrix $L$ in (22) is set to the loadings of the final five principal components of the sample's simulated bond yields. Note that (22) differs from the specification of noise in the data-generating process. Hence estimation maximizes a misspecified likelihood. True ML requires Kalman filter estimation. The computational demands of Kalman filter estimation limit its applicability in Monte Carlo settings. An earlier version of the paper used Kalman filter estimation, but the time required for estimation limits the number of robustness checks that can be performed. The results in the earlier version are very close to those discussed here.

Within a given simulated sample of $T$ time series observations, the accuracy of the no-arbitrage model's cross-sectional estimates for bond $i$ is measured by

$$RMS_i^{noarb} = \left( \frac{1}{T} \sum_{t=1}^{T} \left( Y_{i,t} - \widehat{Y}_{i,t}^{noarb} \right)^2 \right)^{1/2} . \tag{24}$$

This is the root mean squared (RMS) difference between true bond yields, which are not contaminated by measurement error, and fitted yields. Fitted yields are determined by the estimated mapping (5).[5] A similar equation measures the accuracy of the linear factor model OLS mapping. Note that (24) can be calculated using simulated yields but not actual Treasury yields, which are only observed with noise. The disagreement between the two

[5]The reader may ask why model accuracy is not measured by differences between true and fitted parameters of (5). Since the matrix $P$ is simulation-specific, the economic interpretation of these differences is not constant across simulations.

models is measured with

$$RMS_i^{diff} = \left( \frac{1}{T} \sum_{t=1}^{T} \left( \widehat{Y}_{i,t}^{OLS} - \widehat{Y}_{i,t}^{noarb} \right)^2 \right)^{1/2}. \tag{25}$$

These RMS differences are computed for each simulation.

Table 3 reports means and percentiles of the Monte Carlo distributions of these three RMS differences, calculated with 1000 simulations. Because the numbers do not differ much from one bond to another, results are reported for only five of the eight simulated bond yields. There are two clear conclusions to draw from the table. First, the two models have almost identical in-sample cross-sectional accuracy. The no-arbitrage model is slightly more accurate, but on the order of a basis point. Second, the yields implied by the two models' cross-sectional mappings are very similar. Differences between them are seldom greater than five basis points.

A brief discussion of results for the five-year bond is sufficient to justify these conclusions. The interquartile range of the RMS difference between the true yield and the yield fitted by the no-arbitrage model is six to seven basis points. This difference is is created by sampling uncertainty in the coefficients of (3) and by the measurement error in the state vector. For the linear factor model, the corresponding range is seven to nine basis points. Hence imposing no-arbitrage improves the cross-sectional fit by one to two basis points. The interquartile range of the difference between the two models' fitted five-year bond yields is three to five basis points. Moreover, for any other bond listed in the table, the corresponding numbers are even smaller.

### 3.3.2 Comparing out-of-sample forecasts

Because differences between the two cross-sectional mappings are so small, it is not surpring that the two Gaussian models generate similar forecasts of future yields. The simulations discussed here focus on out-of-sample forecasts.

Each simulation consists of $T$ time series observations of the same eight bond yields studied in the simulations of Section 3.3.1. The two Gaussian models are estimated with conditional maximum likelihood using the first $T - 12$ observations. As in the previous simulations, $P$ is the matrix of loadings of the first three principal components of yields in the sample of $T - 12$ observations.

The resulting two models are used to forecast, as of month $T - 12$, the eight yields in months $T - 11$ (one month), $T - 9$ (three months) and $T$ (twelve months). At each horizon there are two competing out-of-sample forecasts. Table 4 summarizes the root mean squared

differences between the two forecasts.

For robustness, results in Table 4 are displayed for three data-generating processes, two sample lengths, and two choices of the standard deviation of measurement error. The first data-generating process is the one described in Section 3.3.1. The second is also a three-factor unrestricted essentially affine model, with parameters determined by estimating the model over the sample October 1980 to September 1989. This sample includes much of the period when the Federal Reserve conducted its monetarist experiment. Of the rolling ten-year samples summarized in Table 2, it has among the highest residual standard errors.[6] Thus this "true" data-generating process is unusual because has dynamics similar to Treasury yield dynamics in the 1980s.

The third data-generating process is a five-factor restricted essentially affine Gaussian model. The functional form is recommended in Duffee (2011). Key features of this model are discussed in Section 4.3. The parameters are determined by estimating the model over the full sample 1972 through 2009. This process is included to study the relative performance of the two estimated models when they are both misspecified.

The most important message of Table 4 is that for all of the reported combinations of data-generating processes, sample sizes, and standard deviations of measurement error, the no-arbitrage and linear factor models produce nearly identical forecasts. The largest number in the entire table is only eight basis points. A couple of other patterns in the table are consistent with our intuition. Disagreements between the forecasts of the two estimated models are larger when the standard deviation of measurement error is higher and when the number of time series observation is smaller. With ten years of data, a standard deviation of ten basis points produces RMS differences in forecasts in the range of one to two basis points. Doubling the standard deviation raises the RMS differences to around two to five basis points. Then cutting the sample in half raises them further to around two to six basis points.

Another pattern might appear puzzling. Disagreements between the two forecasts are smaller at longer forecast horizons. Recall the two models have identical estimated dynamics of the state vector. When the estimated dynamics are stationary (the typical case), forecasts of yields at longer horizons tend to be closer to their unconditional means than forecasts at shorter horizons. This damps the effect of disagreements across the two models in estimated factor loadings.

Finally, consider the role of the data-generating process. Since only three different processes are examined, we cannot draw broad conclusions about the results. The simulations

[6]More precisely, among all of the 337 rolling samples for which the sample physical dynamics (19) are stationary, this period has the largest standard deviation of the residual to the fit of the ten-year bond yield.

suggest that, as long as the number of factors in the estimated models match the number of factors in the true data-generating process, the specific parameters of that true process do not matter much. Here, both estimated models have three factors. Holding the sample size and the standard error of measurement error constant, the two different three-factor data-generating processes produce RMS differences that are all within a basis point of each other. When both estimated models are misspecified because the true process has five factors, the RMS differences are larger. Nonetheless, they remain economically small.

Accuracy of these out-of-sample forecasts is measured by root mean squared forecast error (RMSE) across 1,000 simulations. Table 5 reports the unrestricted essentially affine RMSE less the linear factor model RMSE, in basis points. Negative numbers mean that the no-arbitrage model produces more accurate forecasts. Although the results are important, it is difficult to say much about them other than what is obvious from the table. Differences in forecast accuracy across the two models are economically indistinguishable from zero, regardless of sample size, standard deviation of measurement error, true data-generating process, or forecast horizon. Differences are typically measured in the hundredths of basis points. Every number in the table is less than one basis point in absolute value. There are roughly as many positive values as negative ones.

To summarize, the Monte Carlo evidence shows that when the true data-generating process is Gaussian and satisfies the Duffie-Kan restrictions, an estimated linear three-factor model produces out-of-sample forecasts that are as accurate as forecasts produced by an unrestricted essentially affine three-factor model.

## 3.4 Do Duffie-Kan restrictions hold in Treasury yields?

The intuition behind the irrelevance of cross-sectional restrictions can help us judge whether the cross section of Treasury yields satisfies the Duffie-Kan restrictions for a particular affine model. Here I investigate whether the cross section is consistent with a three-factor unrestricted essentially affine Gaussian model. Under the joint null hypothesis that the model is correct and that we observe yields contaminated by a modest amount of measurement error, the root mean squared disagreements of (25) should be small. In other words, the cross-sectional mapping implied by the no-arbitrage model should be close to that implied by a linear model that does not impose no-arbitrage. Monte Carlo simulations help us judge what "small" means in this context.

I estimate the three-factor unrestricted essentially affine model and its linear factor model counterpart using monthly Treasury yields described in Section 2.5. The estimation methodology is described in Section 3.2. I also estimate the models over rolling 120-month samples

that are studied in Section 2.5. The RMS differences of (25) are reported in Table 6. For the full sample, disagreements in cross-sectional fit range from two to five basis points. The mean disagreements across the rolling samples are smaller, ranging from one to three basis points. Even the largest disagreements across the rolling samples do not exceed six basis points.

From an economic perspective, these disagreements are close to the Monte Carlo results reported in Table 3, for which the true data-generating process is a three-factor unrestricted essentially affine model. But from a hypothesis-testing perspective, the disagreements are too large. I use Monte Carlo simulations to compute distributions of the RMS disagreements for this sample size when the true data-generating process is the one used in Section 3.3.1 is correct. With that process, the measurement error in yields has an unconditional standard deviation of twenty basis points. The table reports, in brackets, 95th percentile values of the RMS statistics. At the shorter end of the yield curve, these bounds are so small (1.3 and 3.4 basis points for three-month and one-year yields, respectively) that we can reject the null, even with the unrealistically large amount of measurement error. With a more plausible ten basis point standard deviation, each of the full-sample RMS disagreements exceed 95 percent bounds. (These bounds are not reported in any table.)

One interpretation of this statistical rejection is that the linear three-factor model is correctly specified, but that owing to market imperfections, the Duffie-Kan restrictions do not hold. This is the argument made in Section 2.2. There is, however, an alternative, equally plausible interpretation. Duffie-Kan restrictions may hold, but they may be the restrictions that apply to a higher-dimensional model. In this case, both of the estimated models that underlie the disagreements in Table 6 are misspecified because there are more than three factors. One sign that these estimated models have too few factors is disagreement between the no-arbitrage and linear factor model.

The last two sets of simulation results in Table 4 illustrate the point. The simulations are generated by a five-factor restricted essentially affine Gaussian model. Yields are contaminated by measurement error with unconditional standard deviations of either ten or twenty basis points. The two three-factor models are then estimated using the simulated data. Root mean squared disagreements between their out-of-sample forecasts are noticeably larger—up to four times as large—as the corresponding disagreements produced by three-factor data-generating processes with equivalent measurement error.

Monte Carlo simulations indicate that the magnitude of the cross-sectional disagreements reported in Table 6 is consistent with misspecifying the number of factors. The data-generating process is the five-factor model of Duffee (2011) used to produce the two sets of simulation results at the bottom of both Table 4 and Table 5. The two three-factor

models are estimated using the simulated data and the cross-sectional disagreements are computed. The table reports, in brackets, 95th percentile values of the RMS statistics (25). The RMS disagreement for the one-year bond is slightly outside of its bound. The other RMS values are comfortably within their bounds.

# 4 Dynamic restrictions

Section 2 shows that the cross-sectional mapping from factors to yields can be estimated with minimal sampling error whether or not Duffie-Kan restrictions are imposed. Therefore when they are imposed—in other words, when equivalent-martingale dynamics are specified—these dynamics are estimated with high precision. Section 3 uses the intuition of Joslin et al. (2011) to explain why this high precision does not help predict future realizations of the state vector in an unrestricted essentially affine model. The main reason is that the physical and equivalent-martingale dynamics share only conditional volatilities. Hence there is no information in the cross section about conditional physical measure expectations.

However, researchers often use the Gaussian essentially affine framework to impose restrictions on the dynamics of risk compensation. These additional restrictions affect forecasts because they link conditional means of the state vector under the two measures. In the context of equation (19), the the vector $\mu_P$ and the matrix $K_P$ are partially determined by the equivalent-martingale parameters (11). As observed by Ball and Torous (1996), such restrictions (if correct) improve the estimation precision of physical dynamics.

This section argues that restrictions on the dynamics of a term structure model can improve out-of-sample forecast accuracy. But the restrictions that are most important empirically do not require a no-arbitrage framework. Instead, I advocate a three-factor dynamic model in which the first principal component of yields follows a random walk, while the second and third principal components follow unrestricted stationary processes. Out-of-sample forecasts of this model dominate those of a wide variety of competing dynamic term structure models.

The first subsection discusses earlier empirical analysis to put this argument in context. The second describes the recommended model and the third describes the horserace and its contestants. Results are in the fourth subsection.

## 4.1 What we know about restrictions on dynamics

In principle, tightening the links between physical and equivalent-martingale dynamics can have large effects on yield forecasts. There is not enough information in just the time series

of Treasury yields to estimate accurately their dynamics. There are two reasons. First, yield dynamics are close, both economically and statistically, to nonstationary. The survey of Martin, Hall, and Pagan (1996) concludes that the level of yields appears to have a unit root, while spreads between yields of different maturities are stationary. Second, researchers are moving to higher-dimension state vectors—typically four or five factors—in response to the evidence in Cochrane and Piazzesi (2005) that small variations in some forward rates can predict large variations in expected excess returns to Treasury bonds. Duffee (2010) finds that when only information from the time series is used to parameterize the physical dynamics of such high-dimensional essentially affine Gaussian models, the resulting estimates wildly overfit in-sample yield behavior.

Unfortunately, it is not clear how to choose a reasonable model of risk compensation dynamics. One approach uses a researcher's economic intuition to impose structure on risk premia. For example, Joslin et al. (2010) assume that the compensation investors require to face interest-rate risk varies with levels of economic activity and inflation. Duffee (2010) restricts conditional Sharpe ratios to a "plausible" range. Another approach rules out any variation in risk premia other than that which is needed to capture what the researcher views as the most important features of risk premia dynamics. Examples include the models of Cochrane and Piazzesi (2008) and Duffee (2011). Both models require that a single factor drive variations in risk premia across all bonds. These views are formed, at least in part, by prior empirical analysis of yield dynamics. An approach tied more explicitly to in-sample yield dynamics is to estimate an unrestricted essentially affine model, then set to zero any parameters for which the estimates are statistically indistinguishable from zero. One example is Duffee (2002).

Another way to increase estimation precision is to impose restrictions directly on term structure dynamics, bypassing restrictions on risk compensation. Diebold and Li (2006) build a dynamic version of the term structure introduced by Nelson and Siegel (1987). Although Christensen et al. (2010) show that the model can be interpreted as a set of restrictions on risk premia dynamics, the restrictions are not motivated by beliefs about risk compensation.

Restrictions on dynamics can be evaluated with out-of-sample tests. Duffee (2002) sets a simple bar for evaluating out-of-sample forecasts from a dynamic term structure model: can it forecast better than the assumption that yields follow a random walk? He shows the entire class of completely affine term structure models, as characterized by Dai and Singleton (2000), fail to clear the bar. He also finds that a three-factor essentially affine Gaussian model clears the bar. This evidence is based on forecasts for 1995 through 1998. Diebold and Li (2006) conclude their model not only clears the bar, but is also more accurate than the three-factor Gaussian model of Duffee (2002). Their out-of-sample forecasts are

produced for January 1994 through December 2000.

## 4.2 A simple dynamic model that ignores no-arbitrage

The focus in this section is on forecasts produced by dynamic term structure models. In the spirit of Duffee (2002), I construct a simple dynamic model that sets a bar for evaluating out-of-sample forecasts. Following Litterman and Scheinkman (1991), the model has three factors, which are the first three principal components of yields. Following Martin et al. (1996), there is a single unit root. Adopting the bar of Duffee (2002), the first principal component of yields, typically referred to as the level of the term structure, follows a random walk.

The parameter restrictions on factor dynamics (19) that produce this model are

$$\mu_P = \begin{pmatrix} 0 & \mu_{P2} & \mu_{P3} \end{pmatrix}', \qquad K_P = \begin{pmatrix} 1 & 0 & 0 \\ 0 & k_{22} & k_{23} \\ 0 & k_{32} & k_{33} \end{pmatrix}. \tag{26}$$

The first row of $K_P$ implies that the first principal component is a random walk. A somewhat more general model that is also explored here replaces the zeros in the first row with free parameters. In this case, the first principal component has a unit root but changes are partially forecastable. The zeros in the first column of $K_P$ are necessary for stationarity of the second and third principal components. The model is completed with the cross-sectional mapping (7). Following the results of Section 2, neither Duffie-Kan nor any other restrictions are placed on the cross-sectional mapping.

The next two subsections determine whether any commonly-used dynamic term structure model can clear the bar set by this model. It is important to recognize that this simple model is built to be consistent with the observed behavior of Treasury yields through 1998, which is the end of the sample used by Duffee (2002). Thus we are especially interested in how this model performs over a more recent period.

## 4.3 A broad look at forecast accuracy

We now run a horse race with ten competing dynamic term structure models. Although the assumption that all yields follow random walks is not a dynamic term structure model, it is included in the horse race as a useful benchmark. Because the objective of this paper is to compare alternative dynamic term structure models, no forecasting regressions included in the horse races. The dynamic models are listed below.

- General linear factor Gaussian models

  Three-, four-, and five-factor versions of the Gaussian linear factor model are estimated. Denote the models LF3, LF4, and LF5. The relevant equations are (19) and (7) with the constraint (8). These models, which do not impose no-arbitrage, have 38, 50, and 63 free parameters respectively. The estimation methodology follows Section 3.2, where the $P$ matrix contains the loadings of the first $n$ principal components of the sample's bond yields.

- Unrestricted essentially affine Gaussian models

  No-arbitrage versions of LF3, LF4, and LF5 are also estimated. Denote the models UEA3, UEA4, and UEA5. The relevant equations are (9), (10), and (19), where (20) links the two measures. The models have 22, 35, and 51 free parameters respectively. The estimation methodology follows Section 3.2.

- The model of Diebold and Li (2006)

  Diebold and Li estimate their model both in unrestricted form, which allows VAR(1) dynamics for level, slope, and curvature, and in restricted form, imposing AR(1) dynamics on each factor. They advocate the latter for out-of-sample forecasting, and I use this restricted form here. Denote the model DL2006. Following Diebold and Li, estimation uses the Kalman filter. The model has 13 free parameters, including the standard deviation of measurement error.

- The five-factor Gaussian model of Duffee (2011)

  This model tightly restricts risk premia dynamics. Only the first two principal components of shocks to the term structure have nonzero prices of risk. Only the first of these has a time-varying price of risk. Denote this model DU2011. Following Duffee, estimation uses the Kalman filter. The model has 29 free parameters, including the standard deviation of measurement error.

- Restricted linear three-factor Gaussian models

  One of these two models is the three-factor Gaussian model with restrictive dynamics (26). Denote this model by PC-RW, for "principal component random walk." The other is the unit-root generalization of (26), where the zeros in the top row of $K_P$ are replaced with free parameters. Denote this model by PC-UR. In both cases, the $P$ matrix is the loading of the first three principal components of the sample's bond

yields. Estimation is with conditional maximum likelihood, assuming that (22) describes measurement error in yields. The PC-RW model has 32 parameters and the PC-UR model has 34 parameters.

The Treasury yield data are described in Section 2.5. Each model is estimated on rolling samples of $T$ months, then used to predict Treasury yields at $T + 3$ and $T + 12$ months. Forecast accuracy is measured by root mean squared forecast errors (RMSE), but not the RMSE's of individual bond yields. Yield forecast errors are highly correlated across bonds, thus there is not much independent information across them. Instead, RMSEs are reported for the the five-year yield (a proxy for the level of the term structure), the five-year yield less the three-month yield (slope), and the two-year yield less the average of the three-month and five-year yields (curvature).

No standard errors are computed. In a data sample mined as extensively as Treasury yields, hypothesis tests for out-of-sample statistics mean little. Therefore the questions we ask and the conclusions we draw are qualitative rather than quantitative. In particular, we want to understand why the simple model of Section 4.2 works relatively well. Is data-mining the reason? Or do regime shifts wreak havoc with forecasts of the other models? Is the main problem with the other models the well-known downward bias in estimates of persistence when persistence is high?

One way to help answer these questions is to compare forecasts produced with models estimated over different sample sizes. In principle, expanding the sample used to estimate term structure models has two competing effects on forecast performance. If the model is specified correctly, expansion raises estimation precision and improves forecast accuracy, especially if yields are highly persistent. But with parameter instability, including more-distant data in estimation of a model that does not allow for instability can reduce forecast accuracy. I therefore use both ten-year ($T = 120$) and twenty-year ($T = 240$) samples. For $T = 120$, there are 325 overlapping rolling samples. The first is January 1972 through December 1981, used to predict yields in March 1982 and December 1982. The last is January 1999 through December 2008, used to predict yields in March 2009 and December 2009. For $T = 240$, there are 205 overlapping rolling samples, with the first predictions made as of December 1991.

Based on these two sample sizes, results are reported for two periods. The first is 1982 through 2009 using $T = 120$. The second is 1992 through 2009, using both $T = 120$ and $T = 240$. Most of the latter sample postdates the sample used by Duffee (2002).[7] It also

[7] An on-line appendix contains a table that reports results for the 1999 through 2009 period, which entirely postdates Duffee's sample. Nothing in it alters the conclusions drawn here.

excludes the period when the Fed's monetarist experiment was winding down as inflation was choked off.

## 4.4 Results

Table 7 reports results for forecasts from 1982 through 2009. Table 8 reports results for forecasts from 1992 through 2009. To avoid getting lost in details, I highlight here the three main conclusions drawn from this exercise. First, the PC-RW model dominates all other models in out-of-sample forecasts. Second, neither the choice to impose no-arbitrage nor the choice of the number of factors matters much when choosing among the linear factor models and the unrestricted essentially affine models. Their forecast accuracies are nearly indistinguishable. Third, the performances of DL2006 and DU2011 models relative to other models critically depend on the period studied.

### 4.4.1 Forecasting the level of the term structure

Begin with the "Level" columns in Table 7, which reports RMSEs for forecasts of the five-year yield during the 1982–2009 period. The most obvious point to take from the results is that differences among the models are not large. At the three-month horizon, the largest difference in RMSEs across all models and is only five basis points. At the twelve-month horizon it is 18 basis points. The discussion here focuses on the longer horizon. Results for the shorter horizon are qualitatively the same and quantitatively muted.

At the twelve-month horizon, the RMSEs for the LF3 and LF4 models and their unrestricted essentially affine counterparts UEA3 and UEA4 are all 147 basis points. The RMSEs for the five-factor versions of these models only slightly larger; 149 and 150 basis points respectively. The surprising aspect of these nearly-identical RMSEs is that the number of factors does not matter much. We know from Duffee (2010) that in-sample, four-factor and five-factor models substantially overpredict variations in yields. Such overfitting suggests that these models would perform relatively poorly out of sample. But over the 1982 through 2009 period there is precisely one epsisode of out-of-sample overfitting by the five-factor models. At the end of October 2008, these five-factor models (which, like all the other models here, are Gaussian), predicted that the three-month yield in October 2009 would be about negative nine percent. If the final three months of the sample are excluded, the LF5 and UEA5 models have RMSEs of 146 and 147 basis points respectively. (These results are not reported in any table.)

All of these models appear to underestimate the persistence in the level of yields. The first row of Table 7 tells us that the simple random walk forecast of yields has an RMSE

15 basis points smaller than those ofthe linear factor models and the unrestricted essentially affine models. Thus not surprisingly, PC-RW has the lowest RMSE among all of the dynamic models. Its RMSE differs slightly from that of the simple random walk forecast the model imposes a random walk on the first principal component instead of on the five-year yield. The more general model PC-UR has a slightly larger RMSE than the simple random walk forecast.

We can use Table 8 to tell similar stories about forecast accuracy in the 1992 to 2009 period. When ten years of data are used in estimation, the general linear factor models and the unrestricted factor models have "Level" RMSEs at the twelve-month horizon in the range of 131 to 136 basis points. The random walk assumption has an RMSE of only 111 basis points. The PC-RW model has an RMSE of 110 basis points.

Do the models other than PC-RW perform poorly because of parameter instability or high persistence in yields? Table 8 tells us, at least for the models here, that high persistence is the culprit. When twenty years of data are used to estimate the general linear factor models, the RMSEs are around 10 basis points smaller than they are when only ten years of data are used. (Only results for the linear factor models are displayed. Results for the unrestricted essentially affine models are almost identical.) This reduction suggests that including more data produces more precise estimates of mean reversion. In the case of the LF5 and UEA5 models, it also eliminates overfitting during the financial crisis. Although using non-yield data is outside the scope of this paper, Kim and Orphanides (2005) note that integrating survey data into model estimation can help estimation precision.

The forecast accuracies of the DL2006 and DU2011 models are mixed. Table 7 reports that over the 1982 through 2009 period, their RMSEs for the five-year yield are close to each other and slightly below those of the models that do not impose random walk or unit root constraints. In Table 8, the same statement holds for models estimated using ten years of data. But when estimating models with twenty years of data, both the DL2006 and DU2011 models have relatively high RMSEs. The only clear result is that neither model outperforms a random walk.

The superiority of the random walk model documented here runs counter to the evidence of Duffee (2002) and Diebold and Li (2006). However, the out-of-sample periods studied in this earlier work are quite short. The data used in the current paper extends their samples by twelve and ten years respectively. In results not detailed in any table, I confirm that these data account for the reversed conclusion. I repeated the empirical analysis of both papers to confirm their results.[8] I then extended their sample periods using more recent

[8]More precisely, I used expanding samples rather than the rolling samples used to construct Tables 7 and 8. I also used the starting points of their data samples, which are 1952 and 1985 respectively.

data and verified that the random walk assumption produced lower RMSEs than either model's forecasts of long-horizon yields.

### 4.4.2 Forecasting slope and curvature

Now consider the "Slope" and "Curve" columns of Tables 7 and 8. Absolute forecast errors are smaller for slope and curvature than they are for the level of the term structure, especially at the twelve-month horizon. The primary reasons are that both the magnitude of the monthly shocks and the persistence of the shocks are smaller for slope and curvature than they are for the level of the term structure.

Since the mean reversion in slope and curvature is not captured by the assumption that yields follow random walks, the random walk forecasts have RMSEs above those of the dynamic term structure models. For example, in Table 7 the random walk forecasts of the slope at the twelve-month horizon have an RMSE of 109 basis points. The corresponding RMSEs for the linear factor models and the essentially affine models are about 10 to 15 basis points smaller. The curvature RMSEs follow a similar pattern. These results carry over to shorter period examined in Table 8.

The only other result that holds across both periods and both estimation sample sizes is that the DL2006 model has slope RMSEs above those of the models with unconstrained dynamics. The performance of the other models is a little scattershot. For example, the DU2011 model has a slope RMSE that is either middle of the road (Table 7), very low (Table 8, $T = 120$), or high relative to other estimated models (Table 8, $T = 240$). The PC-RW model has a relatively high slope RMSE over the entire period 1982 through 2009 (Table 7), but has the lowest slope RMSE over the 1992 through 2009 period (Table 8, both $T = 120$ and $T = 240$).

In summary, the PC-RW model is the clearly preferred dynamic term structure model from the perspective of RMSE. Over the 1992 through 2009 period this model has the lowest RMSE for level and slope, and its RMSE for curvature is equal to the lowest among the dynamic models (for $T = 240$), or only three basis points above the lowest (for $T = 120$). Over the entire 1982 through 2009 period, the RMSE for level is between 16 and 21 basis points lower than the corresponding RMSEs for all other dynamic models. The model's level and slope RMSEs are slightly high, but not enough to offset the model's advantage in forecasting the level of the term structure.

# 5 Conclusion

Dynamic no-arbitrage term structure models have long been recognized as powerful tools for cross-sectional asset pricing. For example, they allow us to price exotic term structure instruments given the properties of standard instruments. But they have nothing special to offer when we are interested in inferring the cross-sectional relation among yields on bonds of different maturities. The reason is that the Treasury cross section almost exactly fits a linear factor model. Cross-sectional relations among yields in a linear factor model are easy to infer from yields without imposing *a priori* restrictions, whether the restrictions are those of no-arbitrage or some alternative model.

By contrast, restrictions on dynamics can improve forecast accuracy. But the most important restriction from an empirical perspective is not one derived from no-arbitrage. Instead, it is the assumption that the level of the term structure, as measured by the first principal component of yields, follows a random walk. A Gaussian three-factor model that satisfies this restriction and is otherwise unconstrained produces out-of-sample forecasts that are more accurate than forecasts produced by the other dynamic models studied here.

# Appendix: Duffie-Kan formulas for the Gaussian model

General Gaussian equivalent-martingale dynamics of the state can be written as

$$x_{t+1} = \mu_X^q + K_X^q x_t + \Sigma_X \epsilon_{t+1}^q$$

where the short rate is

$$r_t = \delta_{0X} + \delta_{1X}' x_t.$$

Denote the log price on an $m$-maturity zero-coupon bond by $p_t^{(m)}$. Applying the intuition of Duffie and Kan (1996), Ang and Piazzesi (2003) show that log bond prices are affine in the state vector. Write the log bond price as

$$p_t^{(m)} = A_m + \mathcal{B}_m' x_t.$$

The loading of the log price on the state vector is

$$B_m' = -\delta_{1X}' \left(I - K_X^q\right)^{-1} \left(I - (K_X^q)^m\right)$$

and the constant term satisfies the difference equation

$$A_1 = -\delta_{0X}, \quad A_{m+1} = -\delta_{0X} + A_m + B_m' \mu_X^q + \frac{1}{2} B_m' \Sigma_X \Sigma_X' B_m.$$

Bond yields are calculated using

$$y_t^{(m)} = -\frac{1}{m} p_t^{(m)}.$$

# References


Ang, Andrew, and Monika Piazzesi, 2003, A no-arbitrage vector autoregression of term structure dynamics with macroeconomic and latent variables, *Journal of Monetary Economics* 50, 745-787.

Backus, David K., and Stanley E. Zin, 1994, Reverse Engineering the yield curve, NBER Working Paper 4676.

Ball, Clifford A., and Walter Torous, 1996, On unit roots and the estimation of interest rate dynamics, *Journal of Empirical Finance* 3, 215-238.

Bekaert, Geert, Robert J. Hodrick, and David A. Marshall, 1997, On biases in tests of the expectations hypothesis of the term structure of interest rates, *Journal of Financial Economics* 44, 309-348.

Christensen, Jens H.E., Francis X. Diebold, and Glenn D. Rudebusch, 2010, The affine arbitrage-free class of Nelson-Siegel term structure models, *Journal of Econometrics*, forthcoming.

Cochrane, John H., and Monika Piazzesi, 2005, Bond risk premia, *American Economic Review* 95, 138-160.

Cochrane, John H., and Monika Piazzesi, 2008, Decomposing the yield curve, Working paper, Chicago Booth.

Cox, John C., Jonathan E. Ingersoll, Jr., and Stephen A. Ross, 1985, A theory of the term structure of interest rates, *Econometrica* 53, 385–407.

Dai, Qiang, and Kenneth J. Singleton, 2000, Specification analysis of affine term structure models, *Journal of Finance* 55, 1943-1978.

Dai, Qiang, and Kenneth J. Singleton, 2002, Expectation puzzles, time-varying risk premia, and affine models of the term structure, *Journal of Financial Economics* 63, 415-441.

Diebold, Francis X., and Canlin Li, 2006, Forecasting the term structure of government bond yields, *Journal of Econometrics* 130, 337-364.

Duffee, Gregory R., 1996, Idiosyncratic variation of Treasury bill yields, *Journal of Finance* 51, 527-552.

Duffee, Gregory R., 2002, Term premia and interest rate forecasts in affine models, *Journal of Finance* 57, 405-443.

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

Duffee, Gregory R., 2011, Information in (and not in) the term structure, *Review of Financial Studies*, forthcoming.

Duffie, Darrell, and Rui Kan, 1996, A yield-factor model of interest rates, *Mathematical Finance* 6, 379-406.

Greenwood, Robin, and Dimitri Vayanos, 2010, Bond supply and excess bond returns, Working paper, LSE.

Joslin, Scott, Marcel Priebsch, and Kenneth J. Singleton, 2010, Risk premium accounting in dynamic term structure models with unspanned macro risks, Working paper, Stanford GSB.

Joslin, Scott, Kenneth J. Singleton, and Haoxiang Zhu, 2011, A new perspective on Gaussian dynamic term structure models, *Review of Financial Studies*, forthcoming.

Kim, Don H., and Athanasios Orphpanides, 2005, Term structure estimation with survey data on interest rate forecasts, FEDS working paper 2005-48, Federal Reserve Board.

Krishnamurthy, Arvind, 2002, The bond/old bond spread, *Journal of Financial Economics* 66, 463-506.

Krishnamurthy, Arvind, and Annette Vissing-Jorgensen, 2010, The aggregate demand for Treasury debt, Working paper, Northwestern University.

Le, Ahn, Kenneth Singleton, and Qiang Dai, 2010, Discrete-time dynamic term structure models with generalized market prices of risk, *Review of Financial Studies* 23, 2184-2227.

Litterman, Robert, and Jose Scheinkman, 1991, Common factors affecting bond returns, *Journal of Fixed Income* 1, 54-61.

Martin, Vance, Anthony D. Hall, and Adrian R. Pagan, 1996, Modelling the term structure, *Handbook of Statistics* 14, G.S. Maddala and C.R. Rao, Eds., 91-118.

Nelson, Charles R., and Andrew F. Siegel, 1987, Parsimonious modeling of yield curves, *Journal of Business* 60, 473-489.

Park, Sang Yong, and Marc R. Reinganum, 1986, the puzzling price behavior of Treasury bills that mature at the turn of calendar months, *Journal of Financial Economics* 16, 267-283.

Vasicek, Oldrich, 1977, An equilibrium characterization of the term structure, *Journal of Financial Economics* 5, 177-188.

Table 1. Cross-sectional fit of a three-factor description of yields

Principal components of a panel of constant-maturity Treasury zero-coupon bond yields are constructed using monthly data from January 1972 through December 2009. The maturities are three months, one through five years, ten years, and fifteen years. Yields on the individual bonds are then regressed on the first three principal components. Standard errors are adjusted for generalized heteroskedasticity and serially correlated residuals using 18 Newey-West lags. The standard deviation of the residuals is $\sigma$ and the serial correlation of residuals at lag $k$ is $\rho(k)$. Yields are in annualized percentage points. All but the three-month yield are constructed by interpolating coupon yields.

| Maturity | Constant | 1st PC | 2nd PC | 3rd PC | $R^2$ | $\sigma$ | $\rho(1)$ | $\rho(19)$ |
|---|---|---|---|---|---|---|---|---|
| 3 mon | −0.138 | 0.385 | 0.584 | −0.605 | 0.9996 | 0.065 | 0.65 | −0.09 |
| | (0.0189) | (0.0011) | (0.0051) | (0.0176) | | | | |
| 1 year | 0.278 | 0.389 | 0.376 | 0.138 | 0.9984 | 0.124 | 0.60 | −0.07 |
| | (0.0365) | (0.0020) | (0.0089) | (0.0329) | | | | |
| 2 years | 0.033 | 0.381 | 0.124 | 0.328 | 0.9995 | 0.070 | 0.49 | −0.02 |
| | (0.0276) | (0.0009) | (0.0065) | (0.0173) | | | | |
| 3 years | −0.035 | 0.367 | −0.039 | 0.337 | 0.9996 | 0.060 | 0.31 | −0.06 |
| | (0.0215) | (0.0008) | (0.0038) | (0.0084) | | | | |
| 4 years | −0.114 | 0.356 | −0.152 | 0.255 | 0.9991 | 0.086 | 0.54 | −0.05 |
| | (0.0277) | (0.0015) | (0.0072) | (0.0221) | | | | |
| 5 years | −0.140 | 0.345 | −0.229 | 0.166 | 0.9992 | 0.079 | 0.66 | −0.04 |
| | (0.0338) | (0.0013) | (0.0069) | (0.0180) | | | | |
| 10 years | 0.020 | 0.304 | −0.425 | −0.265 | 0.9996 | 0.050 | 0.76 | 0.05 |
| | (0.0237) | (0.0009) | (0.0041) | (0.0126) | | | | |
| 15 years | 0.126 | 0.286 | −0.495 | −0.480 | 0.9992 | 0.067 | 0.68 | 0.08 |
| | (0.0283) | (0.0010) | (0.0059) | (0.0171) | | | | |

Table 2. Rolling samples of a three-factor description of yields

This table summarizes cross-sectional regressions of Treasury yields on the three principal components of the Treasury yield term structure. The zero-coupon bonds have maturities of three months, one through five years, ten years, and fifteen years. All but the three-month yield are constructed by interpolating coupon yields. For a given data sample, principal components are defined using the sample covariance matrix of the eight yields. Within the same sample, individual bond yields are then regressed on the first three components. The data are monthly from January 1972 through December 2009. Regressions are estimated on 337 overlapping, rolling samples of 120 months. This table reports the mean and minimum of the 337 $R^2s$ for five of the bonds. It also reports the mean and maximum of the standard deviations of the fitted residuals across these regressions. Finally, it reports the mean and maximum of the serial correlations of residuals.

Panel A. $R^2$s

| | Three months | One year | Five years | Ten years | Fifteen years |
|---|---|---|---|---|---|
| Mean of rolling samples | 0.9996 | 0.9978 | 0.9981 | 0.9991 | 0.9968 |
| Min of rolling samples | 0.9985 | 0.9942 | 0.9966 | 0.9936 | 0.9937 |

Panel B. Standard deviation of residuals (annualized b.p.)

| | Three months | One year | Five years | Ten years | Fifteen years |
|---|---|---|---|---|---|
| Mean of rolling samples | 3.70 | 8.78 | 6.78 | 3.78 | 6.98 |
| Max of rolling samples | 5.41 | 12.31 | 10.90 | 6.07 | 11.23 |

Panel C. Serial correlation of residuals

| | Three months | One year | Five years | Ten years | Fifteen years |
|---|---|---|---|---|---|
| Mean of rolling samples | 0.59 | 0.54 | 0.60 | 0.62 | 0.70 |
| Max of rolling samples | 0.76 | 0.72 | 0.74 | 0.84 | 0.81 |

Table 3. Monte Carlo simulations of cross-sectional mappings from factors to yields

A single Monte Carlo simulation randomly generates ten years of month-end yields on eight zero-coupon bonds. The data-generating process is a three-factor Gaussian model that satisfies no-arbitrage, with parameters given by maximum likelihood estimation over the sample 1972:1 through 2009:12. Observed yields are contaminated by serially correlated measurement error ($\rho = 0.7$) with unconditional standard deviations of 20 annualized basis points.

Two models are estimated using the simulated data. One imposes no-arbitrage. The other uses unrestricted regressions to estimate the cross section. The accuracy of the cross-sectional fits is evaluated using the in-sample root mean squared (RMS) differences between true yields (uncontaminated by measurement error) and model-implied yields. Disagreements between the two models are measured using the in-sample root mean squared differences between the implied yields of the two models.

The table reports means and percentiles of the distributions of these root mean squared statistics across 1000 Monte Carlo simulations. All values are measured in annualized basis points.

| | | Bond maturity | | | | |
|---|---|---|---|---|---|---|
| | Statistic | 3 mo | 1 yr | 5 yr | 10 yr | 15 yr |
| True minus no-arbitrage | Mean | 13.6 | 8.1 | 6.7 | 8.9 | 10.9 |
| | $25^{th}$ | 12.6 | 7.4 | 6.1 | 8.1 | 9.9 |
| | $50^{th}$ | 13.5 | 8.1 | 6.6 | 8.8 | 10.8 |
| | $75^{th}$ | 14.9 | 8.8 | 7.2 | 9.5 | 11.8 |
| | $95^{th}$ | 16.2 | 9.9 | 8.1 | 10.7 | 13.5 |
| True minus regression | Mean | 13.6 | 9.1 | 8.0 | 9.7 | 11.1 |
| | $25^{th}$ | 12.5 | 8.1 | 7.0 | 8.7 | 10.0 |
| | $50^{th}$ | 13.5 | 9.0 | 7.8 | 9.6 | 11.0 |
| | $75^{th}$ | 14.5 | 10.0 | 8.9 | 10.5 | 12.0 |
| | $95^{th}$ | 16.1 | 11.7 | 10.8 | 12.1 | 13.8 |
| No-arbitrage minus regression | Mean | 1.6 | 3.8 | 4.3 | 3.9 | 2.5 |
| | $25^{th}$ | 1.1 | 2.8 | 3.1 | 2.9 | 1.7 |
| | $50^{th}$ | 1.5 | 3.8 | 4.2 | 3.8 | 2.4 |
| | $75^{th}$ | 2.0 | 4.8 | 5.4 | 4.8 | 3.1 |
| | $95^{th}$ | 2.8 | 6.2 | 7.2 | 6.4 | 4.4 |

Table 4. Monte Carlo simulation comparisons of out-of-sample forecasts

Length-$T$ panels of month-end yields on eight bonds are randomly generated from Gaussian no-arbitrage term structure models. The initial $T - 12$ observations are used to estimate two three-factor Gaussian term structure models. The first model imposes no-arbitrage restrictions and the second imposes only a linear factor structure. Each model is then used to forecast the eight bond yields in months $T - 11, \ldots, T$. The table reports root mean squared differences between the two sets of out-of-sample forecasts. All values are in basis points of annualized yields.

The Gaussian no-arbitrage data-generating processes are (I) a three-factor model estimated over the sample 1972:1 through 2009:12; (II) a three-factor model estimated over the sample 1980:10 through 1989:9; (III) a five-factor model estimated over the sample 1972:1 through 2009:12. Observed yields are contaminated by serially correlated measurement error ($\rho = 0.7$) with unconditional standard deviation SD(noise).

| True DGP process | SD(noise), ann. b.p. | $T$ | Months ahead | Bond maturity 3 mon | 1 year | 5 years | 10 years | 15 years |
|---|---|---|---|---|---|---|---|---|
| [I] | 10 | 132 | 1 | 0.8 | 2.2 | 2.4 | 2.0 | 1.3 |
| | | | 3 | 0.8 | 2.0 | 2.2 | 1.9 | 1.1 |
| | | | 12 | 0.7 | 1.7 | 1.9 | 1.6 | 0.9 |
| [I] | 20 | 132 | 1 | 1.9 | 4.4 | 4.8 | 4.2 | 2.7 |
| | | | 3 | 1.7 | 4.1 | 4.5 | 3.9 | 2.3 |
| | | | 12 | 1.5 | 3.6 | 3.9 | 3.3 | 1.8 |
| [I] | 20 | 72 | 1 | 2.7 | 6.0 | 6.3 | 6.0 | 4.1 |
| | | | 3 | 2.5 | 5.5 | 5.9 | 5.6 | 3.6 |
| | | | 12 | 2.2 | 5.2 | 5.5 | 5.0 | 3.0 |
| [II] | 20 | 132 | 1 | 1.2 | 3.5 | 4.6 | 4.2 | 3.0 |
| | | | 3 | 1.0 | 3.2 | 4.1 | 3.8 | 2.6 |
| | | | 12 | 0.8 | 2.7 | 3.5 | 3.2 | 1.9 |
| [III] | 10 | 132 | 1 | 3.9 | 6.5 | 4.7 | 4.8 | 3.8 |
| | | | 3 | 3.7 | 6.2 | 4.6 | 4.7 | 3.7 |
| | | | 12 | 3.5 | 5.9 | 4.4 | 4.7 | 3.6 |
| [III] | 20 | 132 | 1 | 4.6 | 8.0 | 6.7 | 6.6 | 4.9 |
| | | | 3 | 4.4 | 7.6 | 6.5 | 6.4 | 4.7 |
| | | | 12 | 4.1 | 7.2 | 6.2 | 6.2 | 4.4 |

Table 5. Monte Carlo simulation comparisons of forecast accuracy

Length-$T$ panels of month-end yields on eight bonds are randomly generated from Gaussian no-arbitrage term structure models. The initial $T - 12$ observations are used to estimate two three-factor Gaussian term structure models. The first model imposes no-arbitrage restrictions and the second imposes only a linear factor structure. Each model is then used to forecast the eight bond yields in months $T-11, \ldots, T$. For each model, root mean squared yield forecast errors, across 1000 simulations, are calculated. This table reports the RMSE forecast error of the no-arbitrage model less the RMS forecast error of the general linear model. All values are in basis points of annualized yields.

The Gaussian no-arbitrage data-generating processes, labeled [I], [II], and [III], are described in Table 4. Observed yields are contaminated by serially correlated measurement error ($\rho = 0.7$) with unconditional standard deviation SD(noise).

| True DGP process | SD(noise), ann. b.p. | $T$ | Months ahead | Bond maturity: 3 mon | 1 year | 5 years | 10 years | 15 years |
|---|---|---|---|---|---|---|---|---|
| [I] | 10 | 132 | 1 | −0.01 | 0.02 | −0.02 | −0.04 | 0.04 |
| | | | 3 | −0.03 | 0.05 | 0.06 | −0.06 | 0.01 |
| | | | 12 | −0.04 | 0.10 | 0.00 | −0.06 | 0.01 |
| [I] | 20 | 132 | 1 | 0.02 | −0.05 | −0.01 | −0.12 | 0.07 |
| | | | 3 | −0.03 | 0.04 | 0.10 | −0.18 | 0.02 |
| | | | 12 | −0.09 | 0.17 | −0.05 | −0.16 | 0.02 |
| [I] | 20 | 72 | 1 | −0.03 | 0.04 | −0.11 | 0.24 | −0.24 |
| | | | 3 | −0.04 | 0.01 | −0.50 | 0.04 | −0.06 |
| | | | 12 | 0.08 | 0.01 | −0.23 | 0.16 | −0.09 |
| [II] | 20 | 132 | 1 | 0.04 | −0.12 | −0.09 | −0.22 | 0.19 |
| | | | 3 | 0.00 | −0.04 | 0.02 | −0.28 | 0.11 |
| | | | 12 | 0.01 | 0.00 | −0.11 | −0.18 | 0.09 |
| [III] | 10 | 132 | 1 | 0.19 | −0.54 | −0.01 | 0.14 | 0.24 |
| | | | 3 | 0.14 | −0.46 | −0.02 | 0.09 | 0.04 |
| | | | 12 | 0.33 | −0.50 | 0.28 | −0.24 | 0.05 |
| [III] | 20 | 132 | 1 | 0.23 | −0.62 | −0.02 | 0.01 | 0.18 |
| | | | 3 | 0.09 | −0.57 | −0.20 | −0.08 | 0.06 |
| | | | 12 | 0.37 | −0.62 | 0.15 | −0.37 | 0.15 |

Table 6. Cross-sectional mappings from factors to yields, 1972 through 2009

This table summarizes differences in cross-sectional mappings from factors to Treasury yields implied by two three-factor Gaussian models. Two term structure models are estimated using monthly data from January 1972 through December 2009. The same models are also estimated on rolling 120-month subsamples of this period.
One model imposes no-arbitrage. The other uses unrestricted regressions to estimate the cross section. Disagreements between the two models are measured using the in-sample root mean squared differences between the implied yields of the two models. The table reports the root mean squared differences for the full sample, as well as means and maximums across the 337 rolling samples. For the full sample, brackets display upper 95th percentiles of the statistics, computed assuming that the no-arbitrage model is correct. Braces display the same percentiles, computed assuming the true model is a five-factor model described in the text. The percentiles are computed using Monte Carlo simulations. All values are measured in annualized basis points.

| | Three months | One year | Five years | Ten years | Fifteen years |
|---|---|---|---|---|---|
| Full sample | 1.7 | 4.5 | 2.2 | 3.3 | 1.8 |
| | [1.3] | [3.4] | [3.7] | [3.4] | [2.1] |
| | {3.1} | {4.4} | {4.2} | {3.5} | {3.1} |
| Mean of rolling samples | 1.0 | 2.9 | 2.8 | 3.2 | 2.1 |
| Max of rolling samples | 2.7 | 5.6 | 4.2 | 5.4 | 3.9 |

Table 7. Root mean squared errors of monthly out-of-sample forecasts, 1982 to 2009

Term structure models are estimated on rolling panels of Treasury yields with 120 monthly observations. The models are defined in the text. The results are used to forecast the level, slope, and curvature of the term structure three and twelve months ahead. The baseline forecast is the assumption that yields at all maturities follow a random walk. This table reports root mean squared forecast errors in basis points of annualized yields. Forecasts are constructed at month-ends 1981:12 through 2008:12, for a total of 325 observations.

| | RMSE (b.p.) Three months ahead | | | RMSE (b.p.) Twelve months ahead | | |
|---|---|---|---|---|---|---|
| Model | Level | Slope | Curve | Level | Slope | Curve |
| Random walk | 67 | 62 | 24 | 132 | 109 | 34 |
| General linear factor models | | | | | | |
| 3 factors | 68 | 59 | 24 | 147 | 98 | 29 |
| 4 factors | 69 | 58 | 23 | 147 | 96 | 28 |
| 5 factors | 69 | 59 | 23 | 149 | 98 | 29 |
| Unrestricted essentially affine models | | | | | | |
| 3 factors | 69 | 59 | 24 | 147 | 99 | 30 |
| 4 factors | 69 | 58 | 23 | 147 | 96 | 28 |
| 5 factors | 70 | 59 | 24 | 150 | 99 | 30 |
| Models with constrained dynamics | | | | | | |
| Diebold-Li | 70 | 58 | 24 | 143 | 101 | 26 |
| Duffee 5-factor | 70 | 59 | 24 | 144 | 99 | 29 |
| 3-factor random walk | 65 | 60 | 23 | 129 | 103 | 30 |
| 3-factor unit root | 65 | 61 | 24 | 135 | 99 | 29 |

Table 8. Root mean squared errors of monthly out-of-sample forecasts, 1992 to 2009

Term structure models are estimated on rolling panels of Treasury yields with either 120 or 240 monthly observations. The models are defined in the text. The results are used to forecast the level, slope, and curvature of the term structure three and twelve months ahead. The baseline forecast is the assumption that yields at all maturities follow a random walk. This table reports root mean squared forecast errors in basis points of annualized yields. Forecasts are constructed at month-ends 1991:12 through 2008:12, for a total of 205 observations.

| | RMSE (b.p.) Three months ahead | | | RMSE (b.p.) Twelve months ahead | | |
|---|---|---|---|---|---|---|
| Model | Level | Slope | Curve | Level | Slope | Curve |
| Random walk | 57 | 52 | 20 | 111 | 111 | 32 |
| General linear factor models, $T = 240$ | | | | | | |
| 3 factors | 61 | 45 | 21 | 131 | 92 | 31 |
| 4 factors | 62 | 45 | 21 | 131 | 91 | 30 |
| 5 factors | 62 | 46 | 21 | 135 | 93 | 30 |
| Unrestricted essentially affine models, $T = 120$ | | | | | | |
| 3 factors | 61 | 46 | 22 | 132 | 94 | 31 |
| 4 factors | 62 | 46 | 21 | 132 | 91 | 30 |
| 5 factors | 62 | 46 | 23 | 136 | 95 | 32 |
| Models with constrained dynamics, $T = 120$ | | | | | | |
| Diebold-Li | 61 | 48 | 22 | 123 | 100 | 26 |
| Duffee 5-factor | 61 | 49 | 24 | 121 | 85 | 30 |
| 3-factor random walk | 56 | 45 | 21 | 110 | 85 | 29 |
| 3-factor unit root | 57 | 46 | 22 | 112 | 89 | 32 |
| General linear factor models, $T = 240$ | | | | | | |
| 3 factors | 59 | 43 | 20 | 118 | 85 | 26 |
| 4 factors | 60 | 43 | 20 | 120 | 85 | 26 |
| 5 factors | 59 | 44 | 19 | 119 | 85 | 26 |
| Models with constrained dynamics, $T = 240$ | | | | | | |
| Diebold-Li | 64 | 49 | 21 | 133 | 97 | 26 |
| Duffee 5-factor | 67 | 44 | 20 | 138 | 89 | 27 |
| 3-factor random walk | 56 | 42 | 20 | 108 | 83 | 26 |
| 3-factor unit root | 56 | 43 | 21 | 108 | 89 | 27 |

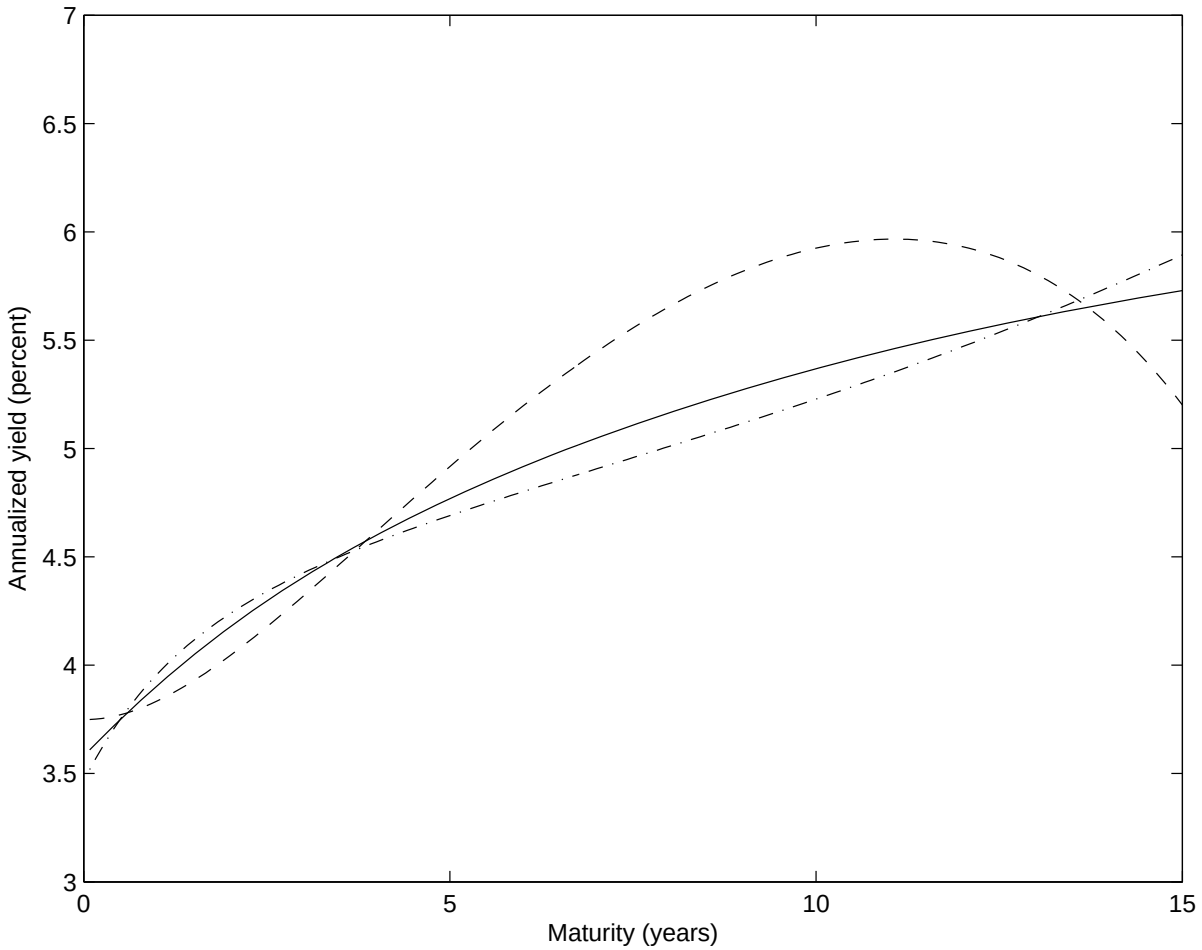


Fig. 1. Hypothetical examples of twelve-month-ahead forecasts produced by no-arbitrage term structure models. Three term structure models have identical specifications of factors, factor dynamics, and forecasts of these factors. The models differ in the no-arbitrage parameters that link bond yields to the factors.

Appendix Table. Root mean squared errors of monthly out-of-sample forecasts, 2001 to 2009

Term structure models are estimated on rolling panels of Treasury yields with either 120 or 240 monthly observations. The models are defined in the text. The results are used to forecast the level, slope, and curvature of the term structure three and twelve months ahead. The baseline forecast is the assumption that yields at all maturities follow a random walk. This table reports root mean squared forecast errors in basis points of annualized yields. Forecasts are constructed at month-ends 2000:12 through 2008:12, for a total of 97 observations.

| | RMSE (b.p.) Three months ahead | | | RMSE (b.p.) Twelve months ahead | | |
|---|---|---|---|---|---|---|
| Model | Level | Slope | Curve | Level | Slope | Curve |
| Random walk | 57 | 59 | 20 | 96 | 120 | 28 |
| General linear factor models, $T = 240$ | | | | | | |
| 3 factors | 63 | 50 | 22 | 132 | 103 | 32 |
| 4 factors | 64 | 50 | 21 | 131 | 103 | 31 |
| 5 factors | 62 | 52 | 22 | 136 | 109 | 32 |
| Unrestricted essentially affine models, $T = 120$ | | | | | | |
| 3 factors | 63 | 52 | 23 | 133 | 105 | 32 |
| 4 factors | 65 | 51 | 21 | 132 | 103 | 31 |
| 5 factors | 63 | 53 | 22 | 136 | 110 | 32 |
| Models with constrained dynamics, $T = 120$ | | | | | | |
| Diebold-Li | 63 | 55 | 20 | 122 | 110 | 25 |
| Duffee 5-factor | 61 | 58 | 27 | 121 | 95 | 32 |
| 3-factor random walk | 56 | 50 | 21 | 100 | 90 | 26 |
| 3-factor unit root | 57 | 53 | 22 | 95 | 99 | 32 |
| General linear factor models, $T = 240$ | | | | | | |
| 3 factors | 58 | 49 | 20 | 107 | 95 | 24 |
| 4 factors | 61 | 49 | 19 | 113 | 95 | 23 |
| 5 factors | 58 | 49 | 20 | 110 | 95 | 24 |
| Models with constrained dynamics, $T = 240$ | | | | | | |
| Diebold-Li | 66 | 55 | 20 | 133 | 111 | 24 |
| Duffee 5-factor | 64 | 50 | 21 | 116 | 94 | 27 |
| 3-factor random walk | 56 | 47 | 20 | 97 | 86 | 24 |
| 3-factor unit root | 56 | 47 | 20 | 95 | 85 | 25 |