A Unified HJM Approach to Non-Markov Gaussian Dynamic Term Structure Models: International Evidence [*]

HAITAO LI[†]
Cheung Kong Graduate School of Business

XIAOXIA YE[‡]
University of Bradford

FAN YU[§]
Claremont McKenna College

July 29, 2016

[*] We are grateful to Areski Cousin, Jin-Chuan Duan, Koji Inui, Masaaki Kijima, Robert Kimmel (discussant), Shaoyu Li, Linlin Niu, Haoxi Yang, and participants at the Young Researchers Workshop in Finance at Tokyo Metropolitan University, the Workshop in Fixed Income and Bond Markets at Xiamen University, and the 9th Annual Risk Management Conference at the National University of Singapore for their helpful comments. The computation in this paper was partially performed on resources provided by the Swedish National Infrastructure for Computing (SNIC) at the Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX). All remaining errors are our own.

[†] htli@ckgsb.edu.cn, address: Oriental Plaza, Tower E3, 3F, One East Chang An Avenue, Beijing, China.

[‡] x.ye4@bradford.ac.uk, address: Emm Ln, Bradford, BD9 4JL, UK.

[§] fyu@cmc.edu, address: Bauer Center, Second Floor, 500 E. Ninth Street, Claremont, CA, USA.

A Unified HJM Approach to Non-Markov Gaussian Dynamic Term Structure Models: International Evidence

Abstract

Motivated by an extensive literature showing that government bond yields exhibit a strong non-Markov property, in the sense that moving averages of long-lagged yields significantly improve the predictability of excess bond returns, we develop a systematic approach of constructing non-Markov Gaussian dynamic term structure models (GDTSMs) under the Heath-Jarrow-Morton (HJM) framework. Compared to the current literature, our approach is more flexible and parsimonious, enabling us to estimate an economically significant non-Markov effect that helps predict excess bond returns both in-sample and out-of-sample across nine industrialized countries.

Keywords: Non-Markov; Gaussian Dynamic Term Structure Models; Excess Bond Returns; International Bond Markets; Moving Averages; Forecasting

JEL classification: C61; E43; E44; G12

I Introduction

The term structure of interest rates is one of the most widely studied topics in economics and finance. Built on the pioneering works of Vasicek (1977) and Cox et al. (1985), a large number of dynamic term structure models (DTSMs) have been developed in the finance literature during the past two decades. According to Dai and Singleton (2003), DTSMs assume that the evolution of the spot rate and the yield curve depends on a finite number of state variables. By judiciously choosing the dynamics of the state variables and their relation with the spot rate, the DTSMs that have been developed to date are empirically flexible and analytically tractable. Among the most prominent classes of DTSMs are the affine term structure models (ATSMs) of Duffie and Kan (1996) and Dai and Singleton (2000), in which the spot rate is a linear function of state variables that follow affine diffusions. These models have been extensively studied in the literature to address a wide range of term structure related issues.

An important feature shared by many DTSMs is that bond yields follow Markov processes. That is, changes in bond yields and excess bond returns depend only on current yields. This is partly attributed to the assumption in these models that the state variables follow either continuous-time or discrete-time Markov processes.[1] However, there is increasing evidence that bond yields are non-Markov. For example, based on nonparametric methods, Chen and Hong (2011) show that the seven-day Eurodollar rates strongly violate the Markov assumption; Cieslak and Povala (2014) find that lagged information (up to a full business cycle) provides additional predictive power for future short rate changes beyond that of the current yield curve alone; Duffee (2011) demonstrates that a single factor "hidden" from current yields can explain almost half of the variation in bond risk premia; Cochrane and Piazzesi (2005) and Feunou and Fontaine (2016) find that, in addition

[1] For example, in the classical setting of Gaussian affine term structure models, the state variables follow continuous-time affine diffusions or discrete-time VAR(1) processes. This, along with a short rate assumed to be linear in the state variables, imply a Markov process for the bond yields.

to the current yields (forward rates), lagged yields explain a significant portion of the variation of bond excess returns.

In light of the above evidence, the extant DTSMs in the literature have been extended to capture non-Markov bond yields, primarily by allowing lagged information to enter the dynamics of the state variables. For example, Ang and Piazzesi (2003) and Jardet et al. (2013) consider macro term structure models, in which the state variables follow a discrete-time Gaussian VAR(p) process under both the physical ($\mathbb{P}$) and risk-neutral ($\mathbb{Q}$) measures. Monfort and Pegoraro (2007) and Gourieroux et al. (2014) develop non-Markov models with finite lags under both $\mathbb{P}$ and $\mathbb{Q}$ within a regime-switching context.[2] Also, using a canonical framework, Joslin et al. (2013) and Feunou and Fontaine (2016) consider non-Markov Gaussian DTSMs, in which the state variables follow a VAR(p) or VARMA(1,1) process under the $\mathbb{P}$ measure. However, both models require the state variables to be VAR(1) (hence Markov) under the $\mathbb{Q}$ measure. With this setup, lagged information only affects the expected change of yields, but has no direct impact on bond pricing. As a result, lagged information is left unspanned by the current yields.[3]

While these non-Markov models provide interesting new insights on term structure dynamics, it is fair to say that non-Markov term structure models are still severely under-studied compared to the large number of Markov DTSMs in the literature. In this paper, we develop a systematic approach of constructing continuous-time non-Markov Gaussian DTSMs under the Heath, Jarrow, and Morton (1992, HJM) framework.[4] Compared to extant DTSMs, HJM models are particularly convenient for modeling the non-Markov features in the data for the following reasons:

First, interest rates generally follow infinite dimensional non-Markov processes under the HJM

[2] In these models, it is still the lags and moving averages of the state variables that generate the non-Markov feature, since the regime-switching behavior is governed by Markov chains with transition probabilities dependent only on the current state. For regime-switching term structure models of the Markov kind, see Dai et al. (2007) and Li et al. (2013).

[3] This is, however, not an issue of consensus in the literature. For example, Feunou and Fontaine (2014) and Bauer and Rudebusch (2015) find evidence supporting the spanning of lagged information by the current yields.

[4] Note that our approach still falls within the general affine framework, and should not be confused with non-affine or nonlinear term structure models, e.g., Ahn and Gao (1999).

framework, and are Markov only under restrictive assumptions. This feature offers great flexibility for modeling bond yields because we can start with a large set of non-Markov models, from which we choose a subset of models that fit the data well. In contrast, the current generation of non-Markov DTSMs build upon Markov DTSMs by extending them in limited ways, such as adding lags to the state variables. If changes in bond yields depend on yields in the distant past, for example, such non-Markov models would need many lags and can become overly complicated to parameterize.[5]

Second, under the HJM framework, the volatility function of the forward rates completely determines the dynamics of the yield curve. For certain volatility functions, the term structure has a Finite-Dimensional Representation (FDR) in the sense that forward rates are linear functions of a finite number of state variables, which follow Gaussian processes. We obtain FDRs of Gaussian HJM models based on the linear systems approach of Björk and Gombani (1999). One important advantage of this approach is that it leads to infinitely many equivalent FDRs for the same HJM model, similar to the "invariant transforms" of Dai and Singleton (2000).[6]

Third, the HJM approach offers potentially richer non-Markov dynamics for the term structure than the current non-Markov DTSMs. Based on appropriate specifications of the volatility function, we can obtain non-Markov Gaussian DTSMs by allowing the number of state variables to be larger than the number of factors, i.e., the dimension of the Wiener process. Specifically, the non-Markov property is present in a subsystem embedded within a larger Markov system in our model. In some sense, this idea is similar to the existing literature based on VAR(p)—a non-Markov process that admits a Markov representation when the state space is expanded to include up to p lags of the state variables. However, the HJM framework allows entire past histories to

[5] Neither VAR(p) nor VARMA(1,1) is flexible enough to be consistent with the empirical data. For example, for three-factor models, VAR(p) becomes intractable when $p > 12$. For monthly data, however, $p = 12$ only accounts for lags up to one year, which does not go far enough to capture the strong non-Markov property exhibited by the data.

[6] Dai and Singleton (2000) use invariant transforms to classify all affine term structure models into a few subclasses of canonical representations.

play a role in the dynamics of the state variables, which can offer greater flexibility than adding a limited number of lags. By imposing additional constraints, we show that our specification reduces to VARMA state variables in discrete time. Moreover, by specifying an appropriate pricing kernel or the market prices of risk as in Joslin et al. (2014), we are able to reproduce the unspanning property of current yields.

We conduct a comprehensive analysis of non-Markov Gaussian term structure models using government bond yields from nine countries, which we divide into three groups based on geographical location or the similarity of economic systems: Asia Pacific (Australia, Japan, and New Zealand), Continental Europe (Germany, Sweden, and Switzerland), and North America and UK (Canada, the United Kingdom, and the United States). We compare the in-sample and out-of-sample performance of the non-Markov models constructed using our approach with that of the non-Markov models in the existing literature, specifically Joslin et al. (2013)'s VAR(4) model and Feunou and Fontaine (2016)'s VARMA(1,1) model. Upon inspecting the in-sample fit to the data using various statistical criteria, such as AIC, BIC, and HIC, we find that our non-Markov models can fit the data in-sample better than both the three-factor Markov model and the non-Markov models of VAR(4) and VARMA (1,1). We also examine the out-of-sample performance of all models in predicting the excess bond returns. The results show that in seven of the nine bond markets, the Markov model fails to capture the excess bond returns out-of-sample in an economically significant way.[7] In contrast, in all countries except for one, there are at least three non-Markov models producing economically significant trading profits. We also find that even though VAR(4) and VARMA(1,1) have similar in-sample trading profits as the continuous-time Markov and non-Markov models, they fail to match the out-of-sample performance of our non-Markov models. Looking at the results across the nine countries, different markets call for different non-Markov

[7] Since we do not consider transaction costs, we require that the associated trading rule to yield a risk-adjusted average return of 30 percent or more to be qualified as being economically significant.

specifications. These observations suggest that the non-Markov property is a staple feature of the bond market. We contribute to the literature by offering a flexible and parsimonious modeling framework that can accommodate this feature.

The remainder of the paper proceeds as follows: In Section II, we review the available empirical evidence on the non-Markov property under the $\mathbb{P}$ and $\mathbb{Q}$ measures, which motivates our theoretical approach. In Section III, we develop a systematic method for building non-Markov GDTSMs under the HJM framework. In Section IV, we conduct an out-of-sample exercise that identifies the non-Markov model that delivers the best performance in capturing the excess bond returns. We conclude with Section V. The proofs and technical details can be found in Appendix A to Appendix D, and we discuss the specification of unspanned risks within our framework in Appendix E.

II Related literature on the non-Markov property

In this section, we first review the extensive empirical evidence on non-Markov state variables under the $\mathbb{P}$ measure. Second, we assess the ongoing debate regarding whether bond yields contain information from lagged risk factors, and we argue that the lack of a consensus makes it necessary to develop term structure models in which the state variables are also non-Markov under the $\mathbb{Q}$ measure.

A Non-Markov property under the $\mathbb{P}$ measure

Taking advantage of the VAR approach, Evans and Marshall (1998, 2007) statistically model bond yields with various maturities along with macro variables. Inspired by these efforts, Ang and Piazzesi (2003) incorporate both macro and latent variables into a VAR representation of arbitrage-free term structure models. Similarly, motivated by the observation that interest rates exhibit a historical dynamics involving lagged values and switching regimes, Monfort and Pegoraro (2007) model state variables using a switching VAR. Although these authors never explicitly mention the non-Markov property under the $\mathbb{P}$ measure, the good performance of the VAR models with more than one lag provides indirect evidence supportive of the non-Markov property. Furthermore, Cieslak and Povala (2014) find that lagged information, spanning the length of a business cycle, improves the econometrician's prediction of future short-rate changes relative to conditioning on the current yield curve alone. This evidence directly supports the non-Markov property under the $\mathbb{P}$ measure.

Cochrane and Piazzesi (2005) find that a single factor can explain the bulk of the variation in expected excess bond returns. They claim that this dominant factor is non-Markov, as the coefficients for predicting this factor from lagged forward rates do not conform to the tent-shaped pattern by which current forward rates predict excess returns. They also find that adding more lags of this factor improves the return-forecasting performance significantly. Extending CP's results, Feunou and Fontaine (2016) forecast excess bond returns using a distributed-lag model that incorporates a moving average (MA) component, which summarizes the information contained in long lags of the forward rates. By adding this MA component, they find that return predictability almost doubles at the one-month and two-month horizons and increases substantially at longer horizons. Their results not only confirm the significance of CP's return-forecasting factor, but also provide evidence of a non-Markov factor structure attributed to the moving average of long-lagged information.

B Non-Markov property under the $\mathbb{Q}$ measure

Within a Gaussian setup, it might seem difficult to generate non-Markov behavior under the $\mathbb{Q}$ measure. For example, consider the standard bond pricing equation:

$$\exp\left(-y_t^{(n)}\right) = \exp\left(-y_t^{(1)}\right)\mathbb{E}_t^{\mathbb{Q}}\left[\exp\left(-y_{t+1}^{(n-1)}\right)\right].$$

Under the assumption that bond yields are linear in conditionally Gaussian state variables, the above implies that:

$$y_t^{(n)} = y_t^{(1)} + \mathbb{E}_t^{\mathbb{Q}}\left[y_{t+1}^{(n-1)}\right] - \frac{1}{2}\mathbb{V}ar_t^{\mathbb{Q}}\left[y_{t+1}^{(n-1)}\right] \Rightarrow \mathbb{E}_t^{\mathbb{Q}}\left[y_{t+1}^{(n-1)}\right] = constant + y_t^{(n)} - y_t^{(1)}. \qquad (1)$$

Eqn. (1) indicates that the information contained in the current yield curve alone should be enough for forecasting yields under the $\mathbb{Q}$ measure. However, this apparently Markov behavior of bond yields under the $\mathbb{Q}$ measure is completely compatible with a non-Markov behavior of the state variables under $\mathbb{Q}$, for example, if some of the state variables summarize the past history of other state variables, implying that bond yields span lagged information.

Whether the current yields span lagged information is somewhat related to the literature of (un)spanned macro risks, since macro variables are typically highly persistent, i.e., the current values of marco variables contain a significant amount of lagged information (see, e.g., Ang and Piazzesi, 2003; Athanasopoulos and Vahid, 2008). On one hand, Joslin et al. (2013) and Joslin et al. (2014) present evidence of unspanned macro risks by projecting bond yields onto current and lagged values of macro risk factors. On the other hand, Feunou and Fontaine (2014) demonstrate that the expectations of macro risk factors, which are closely linked to lagged information, are spanned by bond yields. Also, in a recent attempt to resolve the spanning puzzle in macro-finance term structure models, Bauer and Rudebusch (2015) find evidence that statistically rejects unspannd macro term structure models. Lastly, the empirical evidence in Ang and Piazzesi (2003), Jardet et al. (2013), and Monfort and Pegoraro (2007) shows that the risk-neutral autoregressive matrices of lagged factor values are statistically significant, which confirms the need for non-Markov state variables under the $\mathbb{Q}$ measure.

In light of the aforementioned empirical results, we believe that it is just as important to allow for the construction of DTSMs with non-Markov state variables under the $\mathbb{Q}$ measure as it is under the $\mathbb{P}$ measure. The systematic approach developed in the next section can accommodate the non-Markov property under both measures.

III A systematic approach to non-Markov Gaussian DTSMs

In this section, we develop a systematic approach of constructing non-Markov Gaussian DTSMs under the Heath, Jarrow, and Morton (1992, HJM) framework. We first discuss the advantages of the HJM approach to non-Markov term structure modeling. Then, we describe the linear systems approach of Björk and Gombani (1999), which allows us to construct infinitely many Markov and non-Markov Gaussian term structure models from the same HJM model. Specializing to a deterministic specification of the forward rate volatility function, we present an algorithm for constructing both Markov and non-Markov term structure models. Finally, we show that our non-Markov models provide more flexibility than those in the existing literature. These new models will be implemented in the subsequent empirical analysis.

A The HJM framework

We begin by briefly introducing the HJM framework. Let $f(t,T)$ represent the instantaneous forward rate at time t for a future date $T > t$, which represents the rate that can be contracted at time t for instantaneous risk-free borrowing or lending at time T. Given $f\left(t,T\right)$ for all maturities between t and T, the price at time t of a zero-coupon bond with maturity T can be obtained as:

$$P\left(t,T\right)=\exp\left\{-\int_{t}^{T}f\left(t,s\right)ds\right\}.$$

The spot interest rate at time t is simply $r_t = f(t,t)$.

HJM model term structure dynamics through the stochastic evolution of the forward rates:

$$df\left(t,T\right)=\mu\left(t,T\right)dt+\sigma^{f}\left(t,T\right)dW_{t},$$

where $W_{m\times 1}$ is an m-dimensional Wiener process under the $\mathbb{Q}$ measure, and $\mu\left(t,T\right)$ and $\sigma^{f}\left(t,T\right)_{1\times m}$ are the drift and volatility of the forward rate, respectively. HJM also establish the following no-arbitrage restriction on the drift of the forward rate process:

$$\mu\left(t,T\right)=\sigma^{f}\left(t,T\right)\left(\int_{t}^{T}\sigma^{f}\left(t,s\right)ds\right)^{\mathsf{T}}.$$

Therefore, the volatility function $\sigma^{f}\left(t,T\right)$ completely determines the drift of the forward rate under the $\mathbb{Q}$ measure.

For convenience, we consider the Musiela parameterization, which uses the time to maturity (denoted by x), rather than the time of maturity, to parameterize bonds and forward rates:

Definition 1 *For all $x\geqslant 0$, the forward rate in the Musiela parameterization, $r\left(t,x\right)$, is defined as:*

$$r\left(t,x\right)=f\left(t,t+x\right)\text{ and }P\left(t,T\right)=\exp\left\{-\int_{0}^{T-t}r\left(t,s\right)ds\right\}.$$

Following Brace and Musiela (1994), the standard HJM drift condition can be rewritten as:

$$\begin{aligned} dr\left(t,x\right) &= \mu_r\left(t,x\right)dt + \sigma\left(t,x\right)dW_t, \\ \mu_r\left(t,x\right) &= \frac{\partial}{\partial x}r\left(t,x\right) + \sigma\left(t,x\right)\int_0^x \sigma\left(t,s\right)^{\intercal} ds, \end{aligned}$$

where $W_t = \left[\left[w_i\right]_{i=1}^m\right]^{\intercal}$, $\sigma\left(t,x\right) = \sigma^f\left(t,t+x\right) = \left[\sigma_i\left(t,x\right)\right]_{i=1}^m$, and $\left[\bullet_i\right]_{i=1}^m$ is a compact notation for a row vector $\left[\bullet_1,\bullet_2,\cdots,\bullet_m\right]$. We then have:

$$r\left(t,x\right) = r\left(0,t+x\right) + \Theta\left(t,x\right) + r_0\left(t,x\right), \tag{2}$$

$$\Theta\left(t,x\right) = \int_0^t \sigma\left(s,x+t-s\right)\int_0^{x+t-s}\sigma\left(s,\tau\right)^{\intercal} d\tau ds, \tag{3}$$

$$r_0\left(t,x\right) = \int_0^t \sigma\left(s,x+t-s\right)dW_s, \tag{4}$$

$$dr_0\left(t,x\right) = \frac{\partial r_0\left(t,x\right)}{\partial x}dt + \sigma\left(t,x\right)dW_t, \quad r_0\left(0,x\right) = 0. \tag{5}$$

From (2) to (5), we can see that $\Theta\left(t,x\right) + r_0\left(t,x\right)$ is the time-varying stochastic component of the forward rate $r\left(t,x\right)$. Under our Gaussian setup, $\sigma\left(t,x\right)$ is time-invariant, i.e., $\sigma\left(t,x\right) = \sigma\left(x\right)$, and the non-Markov property is reflected in the drift term of $dr_0\left(t,x\right)$, $\frac{\partial r_0(t,x)}{\partial x} = \int_0^t \frac{\partial\sigma(x+t-s)}{\partial x}dW_s$, which requires integrating over the entire history of the underlying Wiener process.

B Finite-Dimensional Representation of HJM models

Interest rates under HJM models are generally non-Markov and infinite dimensional (see Björk and Gombani, 1999), which makes their empirical implementation difficult. Consequently, a large literature has been developed to identify conditions that allow Finite-Dimensional Representations (FDRs) of HJM models. One widely recognized sufficient condition for HJM models to exhibit FDRs is a time-invariant volatility that is a deterministic function of time to maturity, which satisfies a multi-dimensional linear ODE with constant coefficients (e.g., Björk and Svensson, 2001, Corollary 5.1 or Chiarella and Kwon, 2003, Assumption 1).[8] Since the time and maturity dependent components of the volatility function are separable, we can obtain Markov state variables by integrating over the historical Brownian shocks.

Following the approach of obtaining FDRs of HJM models developed by Björk and Gombani (1999) based on linear systems theory, we start from the following definition:

Definition 2 *A triplet $\{\mathbf{A}, \mathbf{B}, \mathbf{C}(x)\}$, where $\mathbf{A}$ is an $n \times n$ matrix , $\mathbf{B}$ is an $n \times m$ full rank matrix, and $\mathbf{C}\left(x\right)$ is an n-dimensional row-vector function, is called an n-dimensional realization of the system $r_0\left(t, x\right)$ if $r_0\left(t, x\right)$ has the following representation:*

$$r_0\left(t, x\right) = \mathbf{C}\left(x\right) Z_t, \tag{6}$$

$$dZ_t = \mathbf{A} Z_t dt + \mathbf{B} dW_t, \quad Z_0 = \mathbf{0}, \tag{7}$$

where Z_t is an n-dimensional column vector of state variables.

Björk and Gombani (1999) show that for an HJM model to have an FDR as in (6)-(7), the volatility function must be written as:

$$\sigma\left(x\right) = \mathbf{C}(x)\mathbf{B} = \mathbf{C}_0 \exp(\mathbf{A}x)\mathbf{B},$$

where $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}\left(x\right)$ are given in Definition 2, and $\mathbf{C}_0 = \mathbf{C}\left(0\right)$. Also, one can apply invariant

[8] Specifically, the ith component of σ, $\sigma_i\left(x\right)$ (which is n times differentiable with respect to x), satisfies an nth order ODE of the form

$$\frac{d^n}{dx^n}\sigma_i\left(x\right) - \sum_{j=0}^{n-1} \kappa_{ij}\left(x\right) \frac{d^j}{dx^j}\sigma_i\left(x\right) = 0$$

where $\kappa_{ij}\left(x\right)$'s are continuous deterministic functions.

transforms to the triplet $\{\mathbf{A}, \mathbf{B}, \mathbf{C}(x)\}$ to construct a new realization

$$\left\{M\mathbf{A}M^{-1}, M\mathbf{B}, \mathbf{C}(x)\, M^{-1}\right\}$$

given a nonsingular $n \times n$ matrix M. Then:

$$r_0(t,x) = \mathbf{C}(x)\, M^{-1}(MZ_t)\,,$$
$$d(MZ_t) = M\mathbf{A}M^{-1}(MZ_t)\, dt + M\mathbf{B}dW_t,$$

is another FDR for the same HJM model with a new state vector MZ.

Traditional GDTSMs are time-homogeneous. Although the FDR we obtain for $r_0(t,x)$ is time-homogeneous, the first two components of the forward curve in (2), $r(0, t+x)$ and $\Theta(t,x)$, are not. To preserve the time-homogeneity feature, we construct GDTSMs from HJM models by replacing $r(0,t+x) + \Theta(t,x)$ with $\lim_{t\to\infty} r(0,t+x) + \Theta(t,x)$, essentially assuming that the model has evolved from the distant past. De Jong and Santa-Clara (1999) and Trolle and Schwartz (2009) adopt a similar treatment. Details of this derivation are presented in Appendix A and Appendix B.

C Non-Markov and Markov states

An important feature of this approach is that the number of states is allowed to be larger than the number of factors, i.e., the dimension of the Wiener process. In this subsection, using an invariant transform, we show that any model having more states than factors exhibits a *non-Markov property*. We emphasize that by this we mean that some of the states are non-Markov on their own, i.e., their conditional forecast depends on their current as well as lagged values. However, when we examine all states as a whole, they form a Markov system because the requisite lagged information is included as additional state variables.

Let us consider a model with **B** being $n \times m$ and lower trapezoidal, where $n > m$.[9] Partition **B** as follows:

$$\mathbf{B} = \begin{bmatrix} \underbrace{B_1}_{m \times m} \\ \underbrace{B_2}_{(n-m)\times m} \end{bmatrix}_{n \times m} .$$

Since **B** is a full rank and lower trapezoidal matrix, B_1 is a nonsingular and lower triangular matrix by construction. Define a transform matrix M as:

$$M = \begin{bmatrix} \underbrace{\alpha_1 \mathbb{1}}_{m \times m} & \underbrace{\alpha_2 \mathbb{1}}_{m \times (n-m)} \\ \underbrace{-B_2 B_1^{-1}}_{(n-m)\times m} & \underbrace{\mathbb{1}}_{(n-m)\times(n-m)} \end{bmatrix}_{n \times n} ,$$

where α_1 and α_2 are free scalar parameters, and $\mathbb{1}$ is an identity matrix if it is square, and otherwise an identity matrix with a proper number of appended zero rows at the bottom or columns on the right. Then, the last $n - m$ rows of

$$\mathbf{B}_{\text{new}} \equiv M\mathbf{B} = \begin{bmatrix} \underbrace{\alpha_1 B_1 + \alpha_2 B_2 \left(1:m, 1:m\right)}_{m \times m} \\ \underbrace{\mathbf{0}}_{(n-m)\times m} \end{bmatrix}$$

are zeros rows, where $B_2 \left(1:m, 1:m\right)$ is the first m rows of B_2 when $n - m \geq m$ or B_2 appended below with zero rows when $n - m < m$.

Therefore, the last $n - m$ state variables in $Z_{\text{new},t} \equiv MZ_t$ are not directly subject to contemporaneous Gaussian shocks, and are instead exponentially-weighted averages of the first m state variables. In other words, these $n - m$ state variables by design summarize the history of the first m states (see the simple one-factor example following Theorem 1). Therefore, the first m states are

[9] A full rank **B** with n>m can always be transformed into a lower trapezoidal matrix via an invariant transformation.

non-Markov by themselves, in the sense that their drift can depend on both current and lagged values of the first m states. We formalize this observation using the following proposition, a proof of which is provided in Appendix C:

Proposition 1 *For any FDR with $n > m$, there exists at least one m-dimensional subsystem of the FDR that is non-Markov.*

It is worth noting that when $n = m$, all transformations of the m-dimensional FDR are Markov. Therefore, Proposition 1 clarifies the difference between an FDR with $n = m$ and one with $n > m$. In what follows, we will refer to FDRs with $n = m$ as Markov models, and those with $n > m$ as non-Markov models.

D Volatility specification and base realization

One important advantage of using the HJM framework is that the volatility function completely determines model specification under the $\mathbb{Q}$ measure. In this subsection, we introduce a volatility function that guarantees the existence of FDRs of HJM models. The volatility function we consider is a special case of the most general deterministic function that allows FDRs according to Björk and Svensson (2001) and consists mainly of polynomials and exponentials.[10]

Specifically, for an m-factor Gaussian HJM model, the volatility function is

$$\sigma\left(x\right) \equiv \underbrace{\left[\left[1 \quad x \quad \cdots \quad x^{n_i-1}\right] e^{-k_i x}\right]_{i=1}^{I}}_{1\times\sum_{i=1}^{I} n_i} \times \mathbf{\Omega}_{\left(\sum_{i=1}^{I} n_i\right)\times m}, \tag{8}$$

[10] Björk and Svensson (2001) show that the most general deterministic volatility function that allows FDRs of HJM models is the so-called "quasi-exponential" (or QE) function that has the following general form:

$$\sigma_{\mathrm{QE}}\left(x\right) = \sum_{i} e^{\lambda_i x} + \sum_{j} e^{\alpha_j x}\left[p_j\left(x\right)\cos\left(\omega_j x\right) + q_j\left(x\right)\sin\left(\omega_j x\right)\right],$$

where λ_i, α_j, and ω_j are real numbers, and p_j and q_j are polynomials. Moreover, $\sigma_{\mathrm{QE}}\left(x\right)$ can be written as $\mathbf{C}_0 \exp(\mathbf{A}x)\mathbf{B}$.

where n_i is a natural number, $\begin{bmatrix} 1 & x & \dots & x^{n_i-1} \end{bmatrix}$ is a $1 \times n_i$ row vector, k_i is a positive real number with $k_i \leq k_j$ for $i < j$, and $\left[\begin{bmatrix} 1 & x & \dots & x^{n_i-1} \end{bmatrix} e^{-k_i x}\right]_{i=1}^{I}$ is a $1 \times \sum_{i=1}^{I} n_i$ row vector.[11] The matrix $\mathbf{\Omega}$ satisfies the following restrictions:

1. $\mathbf{\Omega}$ is an $\sum_{i=1}^{I} n_i \times m$ full rank matrix with $\sum_{i=1}^{I} n_i \geq m$, i.e., Rank $(\mathbf{\Omega}) = m$.[12]

2. In $\mathbf{\Omega}$, any $\sum_{i=1}^{j} n_i$th row is not a zero row, for $j = 1, 2, \dots, I$.[13]

3. $\mathbf{\Omega}(1, j) + \sum_{k=1}^{I-1} \mathbf{\Omega}\left(1 + \sum_{i=1}^{k} n_i, j\right) \geq 0$, for $j = 1, 2, \dots, m$.[14]

4. $\mathbf{\Omega}$ is set to a lower trapezoidal matrix (a generalization of the lower triangular form for a non-square matrix) for the purpose of identification.[15]

Given the volatility function (8), a base realization is presented in the following theorem, a proof of which is provided in Appendix C.

Theorem 1 *For the HJM volatility function defined in* (8), *one realization triplet* $\{\mathbf{A}, \mathbf{B}, \mathbf{C}(x)\}$ *is:*

$$\mathbf{C}(x) = \left[\begin{bmatrix} 1 & x & \dots & x^{n_i-1} \end{bmatrix} e^{-k_i x}\right]_{i=1}^{I}, \mathbf{B} = \mathbf{\Omega},$$

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_1 & & & \\ & \mathbf{A}_2 & & \\ & & \ddots & \\ & & & \mathbf{A}_I \end{bmatrix}_{n \times n}, \ \mathbf{A}_i = \begin{bmatrix} -k_i & 1 & & & \\ & -k_i & 2 & & \\ & & -k_i & \ddots & \\ & & & \ddots & n_i - 1 \\ & & & & -k_i \end{bmatrix}_{n_i \times n_i},$$

[11] Though the k_i's could take complex values, Joslin et al. (2011) show that a model with two complex conjugate eigenvalues is empirically equivalent to a model with two real eigenvalues equal to the real and imaginary part of the complex eigenvalues. Therefore, without loss of generality, we restrict the k_i's to be real in our specification.

[12] If $m > \sum_{i=1}^{I} n_i$, some of the parameters in $\mathbf{\Omega}$ are unidentifiable, as the m-factor model degenerates to a $\sum_{i=1}^{I} n_i$-factor model.

[13] If it were, then n_i needs to be reduced until this is no longer the case.

[14] This restriction ensures that the volatility of the derived short rate is non-negative.

[15] In our empirical estimation, $\mathbf{\Omega}$ appears only through $\mathbf{\Omega}\mathbf{\Omega}^{\intercal}$. Therefore, only the lower part of $\mathbf{\Omega}$ is identifiable.

where $n = \sum_{i=1}^{I} n_i$, and $\mathbf{A}$ is in block diagonal form with each block given by $\mathbf{A}_i$, whose non-zero elements are indicated above.

Using this base realization, we present a concrete example of a non-Markov model using an invariant transform. Specifically, we consider a one-factor HJM model with its volatility function given by:

$$\sigma(x) = [\Omega_1 + \Omega_2 x]\, e^{-kx}. \tag{9}$$

The base realization of this model is:

$$dZ_t = \begin{bmatrix} -k & 1 \\ 0 & -k \end{bmatrix} Z_t dt + \begin{bmatrix} \Omega_1 \\ \Omega_2 \end{bmatrix} dW_t.$$

Suppose that $0 < k < 1/4$, and we set:

$$M = \begin{bmatrix} \frac{\left(\sqrt{1-4k}-1\right)\Omega_2}{2\Omega_1} & \frac{-\left(\sqrt{1-4k}-1\right)\Omega_1 - 2\Omega_2}{2\Omega_1} \\ -\frac{\Omega_2}{\Omega_1} & 1 \end{bmatrix},$$

then the new state variables $Z_{\text{new},t} = MZ_t$ have the following dynamics:

$$dZ_{\text{new},t} = \begin{bmatrix} \nu & -\nu \\ 1 & -\vartheta \end{bmatrix} Z_{\text{new},t} dt + \begin{bmatrix} -\frac{\Omega_2^2}{\Omega_1} \\ 0 \end{bmatrix} dW_t.$$

where $\nu = \frac{1}{2} - k - \frac{1}{2}\sqrt{1-4k}$ and $\vartheta = k + \frac{1}{2} - \frac{1}{2}\sqrt{1-4k}$. It can be easily shown that the second state variable in $Z_{\text{new},t}$ is an exponentially-weighted average of the history of the first state variable, i.e.,

$$Z_{\text{new},t}(2) = \int_0^t e^{-\vartheta(t-s)} Z_{\text{new},s}(1)\, ds,$$

and the first state variable "extrapolates" from the second:[16]

$$dZ_{\text{new},t}\left(1\right) = \nu\left[Z_{\text{new},t}\left(1\right) - Z_{\text{new},t}\left(2\right)\right]dt - \frac{\Omega_2^2}{\Omega_1}dW_t.$$

Apparently, $Z_{\text{new},t}\left(1\right)$ by itself is non-Markov since its drift contains its own history. However, $Z_{\text{new},t}$ as a whole is a Markov system.

E Concise model notations

For convenience, we establish some concise notations to distinguish the different models within our framework. The previous subsection shows that the specification of the model is completely determined by two entities: the vector $\left[\left[1, x, \cdots, x^{n_i-1}\right]e^{-k_i x}\right]_{i=1}^{I}$ and the matrix $\mathbf{\Omega}$, where the former determines the number of state variables and the latter the number of factors. We use $\mathbf{N} = [n_1, n_2, \cdots, n_I]$, a $1 \times I$ row vector with $n_i - 1$, $i = 1, 2, \ldots, I$, being the highest order in the polynomial associated with k_i, to designate both the number of state variables $\sum_{i=1}^{I} n_i$ and the block dimensions of the base transfer matrix $\mathbf{A}$. We use m to designate the number of factors. For example, a model with $\mathbf{N} = [2, 2, 1]$ and $m = 3$ has 5 state variables and 3 factors, and its base

[16] As documented in Barberis et al. (2015), investors often use extrapolation to form beliefs about future state variables. Though this simple example features extrapolation, it is just as easy to construct an alternative specification in which the first state variable mean-reverts to the second.

realization is:

$$\mathbf{C}\left(x\right) = \begin{bmatrix} e^{-k_1 x} & xe^{-k_1 x} & e^{-k_2 x} & xe^{-k_2 x} & e^{-k_3 x} \end{bmatrix},$$

$$\mathbf{A} = \begin{bmatrix} -k_1 & 1 & 0 & 0 & 0 \\ 0 & -k_1 & 0 & 0 & 0 \\ 0 & 0 & -k_2 & 1 & 0 \\ 0 & 0 & 0 & -k_2 & 0 \\ 0 & 0 & 0 & 0 & -k_3 \end{bmatrix}, \mathbf{B} = \begin{bmatrix} \Omega_1 & 0 & 0 \\ \Omega_2 & \Omega_6 & 0 \\ \Omega_3 & \Omega_7 & \Omega_{10} \\ \Omega_4 & \Omega_8 & \Omega_{11} \\ \Omega_5 & \Omega_9 & \Omega_{12} \end{bmatrix}.$$

F Comparing with existing non-Markov models

The recent term structure literature has shown an increasing interest in non-Markov models. Two studies in this aspect are Joslin et al. (2013) and Feunou and Fontaine (2016). Following Joslin et al. (2013)'s notations, these two models can be summarized as follows:

$$r_t = \rho_0 + \rho_1 Z_t,$$
$$Z_t = K_{0Z}^{\mathbb{Q}} + K_{1Z}^{\mathbb{Q}} Z_{t-\Delta t} + \sqrt{\Sigma_Z}\epsilon_t^{\mathbb{Q}}, \epsilon_t^{\mathbb{Q}} \sim N(0, \mathbb{1}),$$

where r_t is the short-term interest rate, Δt denotes the time interval, and Z_t is a vector of observable states. For example, Z_t can consist of the principal components of bond yields and macro variables.[17]

The difference between these models lies in their specification of the dynamics of Z_t under the $\mathbb{P}$ measure. In Joslin et al. (2013), Z_t is a VAR(p) process under the $\mathbb{P}$ measure:

$$Z_t = K_{0Z}^{\mathbb{P}} + K_{1Z}^{\mathbb{P}} \mathcal{Z}_{t-\Delta t}^{p} + \sqrt{\Sigma_Z}\epsilon_t^{\mathbb{P}}, \epsilon_t^{\mathbb{P}} \sim N(0, \mathbb{1}),$$

[17] Joslin et al. (2013) show that this model is observationally equivalent to a canonical representation of affine models, in which r_t is a linear function of some latent state variables.

where $\mathcal{Z}_{t-\Delta t}^{p} \equiv \left(Z_{t-\Delta t}, \ldots, Z_{t-p\Delta t}\right)$. In Feunou and Fontaine (2016), Z_t is a VARMA(1,1) process under the $\mathbb{P}$ measure:

$$(Z_t - v) = K_{1Z}^{\mathbb{P}} \left(Z_{t-\Delta t} - v\right) + \sqrt{\Sigma_Z}\epsilon_t^{\mathbb{P}} + M_1\sqrt{\Sigma_Z}\epsilon_{t-\Delta t}^{\mathbb{P}}, \epsilon_t^{\mathbb{P}} \sim N(0, \mathbb{1}),$$

where v is the unconditional $\mathbb{P}$-mean of Z_t.

Apparently, under the $\mathbb{Q}$ measure, both Joslin et al. (2013) and Feunou and Fontaine (2016) restrict the state variables to be Markov processes, i.e., VAR(1). Our framework, however, does not impose this restriction. In fact, our approach is more general than the existing non-Markov approach in two aspects: a) we allow the dynamics of the state variables under both $\mathbb{Q}$ and $\mathbb{P}$ to be non-Markov; b) we can incorporate unspanned risks if we restrict the pricing state variables (bond market specific states) to be a proper subset of the entire system of state variables. This unifies the unspanned risk specifications of Joslin et al. (2014), Feunou and Fontaine (2016), and Joslin et al. (2013), and is illustrated in Appendix E.

As shown Feunou and Fontaine (2016), the MA component (over long lags) is more crucial than the AR component (over short lags) in explaining the risk premia. Therefore, the VARMA specification seems to be better than the VAR specification under the discrete-time setting. In the rest of this subsection, we show that our volatility specification can generate VARMA representations.

It is shown in Bergstrom (1983) that the discrete-time representation of a continuous-time vector autoregressive process of order p (CVAR(p)) is precisely a VARMA(p,$p-1$) process. Under our framework, we can easily specify the volatility function to generate a CVAR realization. This is illustrated in the following example, in which we consider a two-dimensional system with $p = 2$.

Following earlier notations, we specify a model with $\mathbf{N} = [1, 1, 1, 1]$, $m = 2$, and the base triplet

$\{\mathbf{A}, \mathbf{B}, \mathbf{C}(x)\}$ given by:

$$\mathbf{C}(x) = \begin{bmatrix} e^{-k_1 x} & e^{-k_2 x} & e^{-k_3 x} & e^{-k_4 x} \end{bmatrix},$$

$$\mathbf{A} = \begin{bmatrix} -k_1 & 0 & 0 & 0 \\ 0 & -k_2 & 0 & 0 \\ 0 & 0 & -k_3 & 0 \\ 0 & 0 & 0 & -k_4 \end{bmatrix}, \mathbf{B} = \begin{bmatrix} \Omega_1 & 0 \\ -\frac{\Omega_1 k_2}{k_1} & 0 \\ \Omega_2 & \Omega_3 \\ -\frac{\Omega_2 k_4}{k_3} & -\frac{\Omega_3 k_4}{k_3} \end{bmatrix}.$$

Considering an invariant transform with:

$$M = \begin{bmatrix} -\frac{1}{k_1} & -\frac{1}{k_2} & 0 & 0 \\ 0 & 0 & -\frac{1}{k_3} & -\frac{1}{k_4} \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix},$$

we have a CVAR(2) realization as:

$$\mathbf{C}_{\text{CVAR}}(x) = \begin{bmatrix} \frac{k_1 k_2 e^{-xk_1} - k_1 k_2 e^{-xk_2}}{k_1 - k_2} & \frac{k_3 k_4 e^{-xk_3} - k_3 k_4 e^{-xk_4}}{k_3 - k_4} & \frac{k_1 e^{-xk_1} - k_2 e^{-xk_2}}{k_1 - k_2} & \frac{k_3 e^{-xk_3} - k_4 e^{-xk_4}}{k_3 - k_4} \end{bmatrix},$$

$$\mathbf{A}_{\text{CVAR}} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -k_1 k_2 & 0 & -k_1 - k_2 & 0 \\ 0 & -k_3 k_4 & 0 & -k_3 - k_4 \end{bmatrix}, \mathbf{B}_{\text{CVAR}} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ \frac{\Omega_1 (k_1 - k_2)}{k_1} & 0 \\ \frac{\Omega_2 (k_3 - k_4)}{k_3} & \frac{\Omega_3 (k_3 - k_4)}{k_3} \end{bmatrix},$$

and denoting the resulting state variables by Z_{CVAR}.

If we normalize the time interval Δt to 1, by Bergstrom (1983, Theorem 2), the first two states in Z_{CVAR} have a VARMA(2,1) representation:

$$Z^{1:2}_{\text{CVAR},t} = F_1 Z^{1:2}_{\text{CVAR},t-1} + F_2 Z^{1:2}_{\text{CVAR},t-2} + \epsilon_t + G\epsilon_{t-1},$$

$$\mathbb{E}\left(\epsilon_t\right) = 0, \;\; \mathbb{E}\left(\epsilon_t \epsilon_t^{\intercal}\right) = K, \;\; \mathbb{E}\left(\epsilon_s \epsilon_t^{\intercal}\right) = 0 \;\; \left(s \neq t\right),$$

where $Z^{1:2}_{\text{CVAR},t}$ denotes the first two states in Z_{CVAR},

$$F_1 = \begin{bmatrix} e^{-k_1} + e^{-k_2} & 0 \\ 0 & e^{-k_3} + e^{-k_4} \end{bmatrix}, F_2 = \begin{bmatrix} -e^{-(k_1+k_2)} & 0 \\ 0 & -e^{-(k_3+k_4)} \end{bmatrix},$$

and K and G satisfy the equations:

$$K + GKG^{\intercal} = \Gamma_0, \;\; GK = \Gamma_1,$$

where

$$\Gamma_0 = \int_0^1 P\left(s\right) \mathbf{B}_{\text{CVAR}} \mathbf{B}^{\intercal}_{\text{CVAR}} P^{\intercal}\left(s\right) + Q\left(s\right) \mathbf{B}_{\text{CVAR}} \mathbf{B}^{\intercal}_{\text{CVAR}} Q^{\intercal}\left(s\right) ds,$$

$$\Gamma_1 = \int_0^1 P\left(s\right) \mathbf{B}_{\text{CVAR}} \mathbf{B}^{\intercal}_{\text{CVAR}} Q^{\intercal}\left(s\right) ds,$$

$$P\left(h\right) = \begin{bmatrix} \frac{e^{-\left(k_1+k_2\right)\left(h+1\right)}\left(e^{k_1+hk_2} - e^{k_2+hk_1}\right)}{k_1-k_2} & 0 \\ 0 & \frac{e^{-\left(k_3+k_4\right)\left(h+1\right)}\left(e^{k_3+hk_4} - e^{k_4+hk_3}\right)}{k_3-k_4} \end{bmatrix},$$

$$Q\left(h\right) = \begin{bmatrix} \frac{e^{-hk_2} - e^{-hk_1}}{k_1-k_2} & 0 \\ 0 & \frac{e^{-hk_4} - e^{-hk_3}}{k_3-k_4} \end{bmatrix}.$$

Under this specification, K and G cannot be solved in closed-form. However, this should not be a concern, since the purpose of this illustration is merely to show that our general framework is capable of generating VARMA representations. Once we have estimated the structural parameters of the triplet, we can solve K and G numerically.

It is worth noting that the volatility function for this VARMA(2,1) representation is constrained. While the lower trapezoidal matrix $\mathbf{B}$ has seven free parameters, we constrain four of these parameters to be functions of other parameters or zero so as to achieve the VARMA representation. If we do not impose these constraints, we will have more flexibility in capturing the non-Markov property under the $\mathbb{Q}$ measure, although the discrete-time state variables will no longer follow a VARMA process. Moreover, the VARMA model presented here should not be directly compared with that of Feunou and Fontaine (2016), as their model is unrestricted under the $\mathbb{P}$ measure and of order (1,1), while the one presented here is restricted under the $\mathbb{Q}$ measure and can be readily extended to higher orders.[18] However, as mentioned in Wymer (1993), a large efficiency gain can be expected of the estimates of the restricted VARMA representation of a continuous-time system relative to the unrestricted VARMA estimates when it comes to prediction.

G Extensions to the traditional affine approach

Although we start from an HJM framework, we cast our approach in a state space formulation to make it comparable to the traditional affine approach. At first sight, our approach, which allows for a non-square matrix $\mathbf{B}$, resembles the traditional affine approach combined with a rank restriction on the covariance matrix. However, this intuition is incorrect as our approach is actually an extension rather than a restriction of the traditional affine approach. To see this, we note in a companion paper that many of the existing GDTSMs are special cases of our framework, in the sense that two seemingly different models can actually originate from very similar HJM volatility functions. Furthermore, when viewed from the traditional affine perspective, e.g., Dai and Singleton (2000) (DS), Collin-Dufresne et al. (2008) (CGJ), and Joslin et al. (2011) (JSZ), our systematic

[18] Under the $\mathbb{P}$ measure, our model will have more degrees of freedom thanks to the market price of risk parameters, although the VARMA representation is usually lost.

approach has several distinctive features.

First, we compare our HJM-based affine framework to DS as the latter is still the most popular setup to specify GDTSM in continuous-time. The canonical parameterization adopted by DS, say for a three-factor model, is given by:

$$\mathbf{C} = \begin{bmatrix} \rho_1 & \rho_2 & \rho_3 \end{bmatrix}, \mathbf{A} = \begin{bmatrix} \kappa_{11} & 0 & 0 \\ \kappa_{21} & \kappa_{22} & 0 \\ \kappa_{31} & \kappa_{23} & \kappa_{33} \end{bmatrix}, \mathbf{B} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Using our notations, this model can be rewritten as:[19]

$$\mathbf{C} = \begin{bmatrix} \Omega_1 + \Omega_2 + \Omega_3 & \Omega_4 + \Omega_5 & \Omega_6 \end{bmatrix},$$

$$\mathbf{A} = \begin{bmatrix} -k_1 & 0 & 0 \\ \frac{\Omega_2}{\Omega_4}(k_1 - k_2) & -k_2 & 0 \\ \frac{\Omega_3\Omega_4(k_1-k_3)-\Omega_2\Omega_5(k_1-k_2)}{\Omega_4\Omega_6} & \frac{\Omega_5}{\Omega_6}(k_2 - k_3) & -k_3 \end{bmatrix}, \mathbf{B} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Under this parameterization, **B** cannot be non-square, because a non-square **B** will cause the system to be unidentified. If we set, for example:

$$\mathbf{B} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix},$$

[19] The base realization of DS following Theorem 1 is:

$$\mathbf{C} = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}, \mathbf{A} = \begin{bmatrix} -k_1 & 0 & 0 \\ 0 & -k_2 & 0 \\ 0 & 0 & -k_3 \end{bmatrix}, \mathbf{B} = \begin{bmatrix} \Omega_1 & 0 & 0 \\ \Omega_2 & \Omega_4 & 0 \\ \Omega_3 & \Omega_5 & \Omega_6 \end{bmatrix}.$$

We obtain the triplet in the main text by transforming **B** into an identity matrix.

then the HJM volatility function $\mathbf{C}\exp\left(\mathbf{A}x\right)\mathbf{B}$ becomes:

$$\begin{bmatrix} \Omega_1 e^{-xk_1} + \Omega_2 e^{-xk_2} + \Omega_3 e^{-xk_3} & \Omega_4 e^{-xk_2} + \Omega_5 e^{-xk_3} \end{bmatrix}.$$

We see that Ω_6 has dropped out of the volatility function, meaning that it cannot be identified. In this respect, only CGJ's parameterization has the same flexibility as our approach on specifying non-square **B**.

Second, DS' parameterization does not accommodate all cases where **A** has repeated eigenvalues. For example, a model with the following base realization according to Theorem 1:

$$\mathbf{C} = \begin{bmatrix} 1 & 1 & 0 \end{bmatrix}, \mathbf{A} = \begin{bmatrix} -k_1 & 0 & 0 \\ 0 & -k_2 & 1 \\ 0 & 0 & -k_2 \end{bmatrix}, \mathbf{B} = \begin{bmatrix} \Omega_1 & 0 & 0 \\ \Omega_2 & \Omega_4 & 0 \\ \Omega_3 & \Omega_5 & \Omega_6 \end{bmatrix}$$

cannot be written in DS' parameterization. To make **B** an identity matrix, **C** and **A** must be

$$\mathbf{C} = \begin{bmatrix} \Omega_1 + \Omega_2 & \Omega_4 & 0 \end{bmatrix}, \mathbf{A} = \begin{bmatrix} -k_1 & 0 & 0 \\ \frac{\Omega_3+\Omega_2(k_1-k_2)}{\Omega_4} & \frac{\Omega_5-\Omega_4 k_2}{\Omega_4} & \frac{\Omega_6}{\Omega_4} \\ \frac{\Omega_3\Omega_4(k_1-k_2)+\Omega_2\Omega_5(k_2-k_1)-\Omega_3\Omega_5}{\Omega_4\Omega_6} & -\frac{\Omega_5^2}{\Omega_4\Omega_6} & -\frac{\Omega_5+\Omega_4 k_2}{\Omega_4} \end{bmatrix}.$$

Apparently, **A** is not a lower triangular matrix if **B** is of full-rank. Neither DS' nor CGJ's parameterization is nearly as convenient as ours in specifying the cases where **A** has repeated eigenvalues. In fact, without linkages between our parameterization and those of DS' nor CGJ's, it is almost impossible to figure out the specification for the cases of repeated eigenvalues under their parameterizations. We show these linkages in the companion paper.

JSZ's canonical parameterization generalizes that of DS in a discrete-time setting by directly specifying a full-rank lower triangular matix **B** and the eigenvalues of **A**.[20] Therefore, they are able to accommodate the cases where **A** has repeated eigenvalues. However, they do not explore the possibility or implication of **B** being non-square or singular. This is probably due to the fact that in discrete-time, the invertibility of **B** is indispensible to a well-defined likelihood function (see equation (19) in JSZ). Therefore, the rank constraint on **B** is not applicable in JSZ. Moreover, specifying only the eigenvalues of **A**, as in JSZ, may lead to ambiguities with respect to the HJM volatility function in our framework.

For example, given the same **B**, it is not clear in JSZ what distinguishes two models with their respective **A** being:

$$\begin{bmatrix} -k_1 & 0 & 0 \\ 0 & -k_2 & 1 \\ 0 & 0 & -k_2 \end{bmatrix} \text{ and } \begin{bmatrix} -k_1 & 0 & 0 \\ 0 & -k_2 & 0 \\ 0 & 0 & -k_2 \end{bmatrix}.$$

In both cases, **A** has a repeated eigenvalue of $-k_2$. However, using our framework, the difference between the two models is revealed when we examine their respective HJM volatility function. When

$$\mathbf{A} = \begin{bmatrix} -k_1 & 0 & 0 \\ 0 & -k_2 & 1 \\ 0 & 0 & -k_2 \end{bmatrix},$$

the volatility function is

$$\begin{bmatrix} \Omega_1 e^{-xk_1} + (\Omega_2 + (1+x)\,\Omega_3)\,e^{-xk_2} & (\Omega_4 + (1+x)\,\Omega_5)\,e^{-xk_2} & (1+x)\,\Omega_6 e^{-xk_2} \end{bmatrix}.$$

[20] In JSZ's notation, our matrix **B** is represented by Σ_X, and the eigenvalues of **A** are represented by $\lambda^{\mathbb{Q}}$.

When

$$\mathbf{A} = \begin{bmatrix} -k_1 & 0 & 0 \\ 0 & -k_2 & 0 \\ 0 & 0 & -k_2 \end{bmatrix},$$

it is

$$\begin{bmatrix} \Omega_1 e^{-xk_1} + (\Omega_2 + \Omega_3)\, e^{-xk_2} & (\Omega_4 + \Omega_5)\, e^{-xk_2} & \Omega_6 e^{-xk_2} \end{bmatrix}.$$

This example highlights the differential impact of the three random sources on the forward curve in the two models, even though they appear indistinguishable in JSZ's eigenvalue specification. The difference can be significant since the first model allows a hump-shaped impact on the forward curve from all three random sources, while the second model does not.

In sum, our HJM-based approach offers two extensions to the traditional affine approach: a) It allows **B** to be a non-square matrix, which opens the door for systematically specifying non-Markov GDTSMs; b) It offers a more general canonical parameterization for **A**.

IV Empirical analysis: An out-of-sample exercise

A Data

The data used in this paper are bond yields from nine industrialized countries with different monetary policies constructed by Wright (2011), and made available through an online appendix.[21] We categorize the nine countries into three groups based on their geographical locations and economic similarities: Asia Pacific (Australia, Japan, and New Zealand), Continental Europe (Switzerland, Sweden, and Germany), and North America and UK (Canada, the United Kingdom, and the United States).

[21] In Wright (2011) 's dataset, zero-coupon yield curves of Norway are also available. However, due to their relatively short history, we exclude Norway's data from our analysis.

The data contain zero yields with maturities ranging from one year to ten years with increments of one year, and observed at the monthly frequency. The sample periods start at different dates for different countries and all end in April 2009. In our analysis, we divide the data into two subsamples. For countries with sample periods starting earlier than January 1980 (DE, GB, and US), the in-sample period ends in April 1994. For the rest of the countries (AU, JP, NZ, CH, SE, and CA), the in-sample period ends in April 2000 or April 2004. The details about the sample periods of different countries are summarized in Table 1.

[Insert Table 1 about here]

B Non-Markov models with a Markov origin

We have shown that any model with **B** being $n \times m$, where $n > m$ (regardless of how **N** is specified), exhibits a non-Markov property. However, in order to take advantage of JSZ's estimation method for Markov GDTSMs, in the empirical analysis we focus on models with not only **B** being $n \times m$, but also **N** being $1 \times m$. Non-Markov models with this type of specifications have a Markov model as their origin. In other words, they all reduce to a Markov model when certain parameters are set to zero.

To see this point more clearly, let us consider again the model with $\mathbf{N} = [2, 2, 1]$ and $m = 3$. If we set Ω_2, Ω_4, Ω_6, Ω_8, Ω_{10}, and Ω_{11} in **B** to zero, i.e.,

$$\mathbf{B} = \begin{bmatrix} \Omega_1 & 0 & 0 \\ 0 & 0 & 0 \\ \Omega_3 & \Omega_7 & 0 \\ 0 & 0 & 0 \\ \Omega_5 & \Omega_9 & \Omega_{12} \end{bmatrix},$$

the model reduces to $\mathbf{N} = [1,1,1], m = 3$, because the volatility functions of $\mathbf{N} = [2,2,1], m = 3$ and $\mathbf{N} = [1,1,1], m = 3$ are exactly the same:

$$\begin{aligned}
&\mathbf{C}_0 \exp(\mathbf{A}x)\mathbf{B} \\
&= \begin{bmatrix} 1 & 0 & 1 & 0 & 1 \end{bmatrix} \exp\left(\begin{bmatrix} -k_1 & 1 & 0 & 0 & 0 \\ 0 & -k_1 & 0 & 0 & 0 \\ 0 & 0 & -k_2 & 1 & 0 \\ 0 & 0 & 0 & -k_2 & 0 \\ 0 & 0 & 0 & 0 & -k_3 \end{bmatrix} x \right) \begin{bmatrix} \Omega_1 & 0 & 0 \\ 0 & 0 & 0 \\ \Omega_3 & \Omega_7 & 0 \\ 0 & 0 & 0 \\ \Omega_5 & \Omega_9 & \Omega_{12} \end{bmatrix} \\
&= \begin{bmatrix} 1 & 1 & 1 \end{bmatrix} \exp\left(\begin{bmatrix} -k_1 & 0 & 0 \\ 0 & -k_2 & 0 \\ 0 & 0 & -k_3 \end{bmatrix} x \right) \begin{bmatrix} \Omega_1 & 0 & 0 \\ \Omega_3 & \Omega_7 & 0 \\ \Omega_5 & \Omega_9 & \Omega_{12} \end{bmatrix}.
\end{aligned}$$

The reason why the model $\mathbf{N} = [2,2,1], m = 3$ has a Markov origin is that its matrix $\mathbf{A}$ is 5×5 but has only three distinct eigenvalues. Thus, by setting some of the Ω_i's to zero, we manage to degenerate $\mathbf{A}$ to 3×3, which will be for a Markov model with the three original eigenvalues. In contrast, the model $\mathbf{N} = [1,1,1,1], m = 3$ does not have a Markov model as its origin because its matrix $\mathbf{A}$ has four distinct eigenvalues and therefore cannot degenerate to a 3×3 matrix.

C Estimation

In this empirical analysis, we consider non-Markov models with four to six state variables, which all have the same three-factor Markov model as their origin. Specifically, using the notations introduced in Section E, $m = 3$ for all these non-Markov models, and their $\mathbf{N}$ vectors are laid out in the third column of Table 2. In total, we will be testing 19 different non-Markov models for each country.

To estimate the model parameters, we need to specify the dynamics of the state variables under the $\mathbb{P}$ measure. Adopting an essentially affine specification of the market price of risk (Duffee, 2002):

$$dW_t^{\mathbb{P}} = (\lambda_1 + \lambda_2 Z_t)\, dt + dW_t,$$

the $\mathbb{P}$-dynamics of Z_t is:

$$dZ_t = \mathbf{A}^{\mathbb{P}} \left(-\mu + Z_t\right) dt + \mathbf{B} dW_t^{\mathbb{P}},$$

where

$$\mathbf{A}^{\mathbb{P}} = \mathbf{A} - \mathbf{B}\lambda_2, \mu = \left(\mathbf{A}^{\mathbb{P}}\right)^{-1} \mathbf{B}\lambda_1.$$

In order to reduce the chances of being stuck in some local optima far away from the global optimum, the parameters of these models are estimated in a two-step procedure. First, the majority of the parameters from the underlying three-factor Markov model are estimated using JSZ's method. Second, using these estimates and zeros for the rest of the parameters as initial values, all parameters are estimated using the Kalman filter in conjunction with ML. These two steps are elaborated in the following two subsections. The ten maturities of the term structure (one to ten years) from the in-sample periods (see Table 1) are used to estimate the parameters.

C.1 Applying JSZ's method to continuous-time Markov GDTSMs

The purpose of applying JSZ's method is not to estimate the $\mathbb{P}$ and $\mathbb{Q}$ parameters separately, but to use the results as initial values to speed up the convergence of the full estimation.

Following JSZ (Case P in their paper), we assume that the first three principal components (PCs) of the yields are observed perfectly, while the yields themselves are observed with error. These measurement errors are assumed to be normally distributed with zero mean and variance

σ^2:

$$\underbrace{\boldsymbol{e}_t}_{10\times 1} \sim N\left(\underbrace{\mathbf{0}}_{10\times 1}, \underbrace{\sigma^2\mathbb{1}}_{10\times 10}\right),$$

and independent across time. Then, the conditional log likelihood function (under $\mathbb{P}$) of the observed yields has two parts:

$$\log ll_{t|t-\Delta t} = \log ll_t^{\text{errors}} + \log ll_{t|t-\Delta t}^{\text{states}}.$$

The first part of the likelihood function represents the log likelihood of measurement errors,

$$\log ll_t^{\text{errors}} \propto -\frac{1}{2}\log\left(\det\left(\sigma^2\mathbb{1}\right)\right) - \frac{\text{error}_t^{\mathsf{T}}\text{error}_t}{2\sigma^2},$$

where

$$\begin{aligned}
\text{error}_t &= y_t - \{C_m + B_m Z_t\}, \\
Z_t &= (wm \cdot B_m)^{-1}\left(\text{PC}_t - wm \cdot C_m\right), \\
C_m &= \underbrace{\left[\varphi + \frac{\int_0^1 \Theta^*(s)ds}{1}, \quad \cdots, \quad \varphi + \frac{\int_0^{10} \Theta^*(s)ds}{10}\right]^{\mathsf{T}}}_{10\times 1}, \\
B_m &= \underbrace{\left[\left(\frac{\int_0^1 \mathbf{C}(s)ds}{1}\right)^{\mathsf{T}}, \quad \cdots, \quad \left(\frac{\int_0^{10} \mathbf{C}(s)ds}{10}\right)^{\mathsf{T}}\right]^{\mathsf{T}}}_{10\times 3}.
\end{aligned}$$

Here, $\underbrace{y_t}_{10\times 1}$ is a vector of zero yields of ten different maturities, and $\underbrace{wm}_{3\times 10}$ is the weighting matrix for the principal components, i.e., $\underbrace{\text{PC}_t}_{3\times 1} = wm \cdot y_t$. The above expression makes use of the relation between the model-implied zero yield and the state variables backed out from the principal components when forward rates are assumed to be time-homogeneous.[22]

[22] See (A5) in Appendix A. The parameter φ and function $\Theta^*(\cdot)$ can also be found therein.

The second part of the likelihood function represents the log transition density of the state variables,

$$\log ll_{t|t-\Delta t}^{\text{states}} \propto -\frac{1}{2}\log\left(\det\left(cv\right)\right) - \frac{\left(\alpha + \beta Z_{t-\Delta t} - Z_t\right)^{\intercal} cv^{-1}\left(\alpha + \beta Z_{t-\Delta t} - Z_t\right)}{2},$$

where $cv = \int_0^{\Delta t} \exp\left(\mathbf{A}^{\mathbb{P}}\left(\Delta t - s\right)\right)\mathbf{B}\mathbf{B}^{\intercal}\exp\left(\mathbf{A}^{\mathbb{P}}\left(\Delta t - s\right)\right)^{\intercal} ds$ is the conditional variance of Z_t given $Z_{t-\Delta t}$, and $\mathbf{A}^{\mathbb{P}}$, α, and β are explicit functions of the data and the $\mathbb{Q}$ parameters, which are determined as follows: First, given the $\mathbb{Q}$ parameters, the state variables are backed out from the principal components. Then, α and β are obtained as the OLS estimates of the following regression:

$$Z_t = \alpha + \beta Z_{t-\Delta t} + \epsilon_t.$$

Finally, we calculate $\mathbf{A}^{\mathbb{P}}$ as $\frac{\log(\beta)}{\Delta t}$, where log denotes the matrix logarithm operator.[23]

Therefore, $\sum_{t=2\Delta t}^{\mathrm{T}\Delta t} \log ll_{t|t-\Delta t}$ is a function of the data and the $\mathbb{Q}$ parameters only. This simplification reduces the number of parameters in the optimization and speeds up the convergence of the estimation. Given the ML estimates of the $\mathbb{Q}$ parameters, $\mathbf{A}^{\mathbb{P}}$ and μ (hence λ_2 and λ_1) are explicitly determined from the values of α and β. By the results of JSZ, such estimates are also ML estimates.

C.2 Estimating the continuous-time non-Markov GDTSMs

It is possible to use the Kalman filter in conjunction with ML to estimate all parameters without any priors. However, recent literature (e.g., Bauer et al., 2012) has found that the log likelihood function of GDTSMs can be badly behaved (very flat in the parameter space), exhibiting many lo-

[23] In all of our estimations, the ML parameter estimates result in β having a unique $\log(\beta)$, i.e., β is nonsingular with no negative eigenvalues, and every eigenvalue of $\log(\beta)$ has an imaginary part lying strictly between $-\pi$ and π. See, e.g., Higham (2008, Theorem 1.31).

cal maxima. Therefore, it is time-consuming to search for the global optimum from uninformative initial values when the number of parameters is large.

To alleviate the difficulty in estimating the non-Markov models, we initialize the parameters of a non-Markov model that have counterparts in its Markov origin (we call them *Markov parameters*) to their estimates using the JSZ method, and the other parameters (we call them *non-Markov parameters*) to zero. Then, all parameters are estimated using the Kalman filter in conjunction with ML. This separation between the Markov and non-Markov parameters is feasible because a) all non-Markov models we consider here have a Markov origin, and b) we use the essentially affine market price of risk specification.

We use the model $\mathbf{N} = [1,1,2]\,, m = 3$ as an example to illustrate this separation. To conserve space, we denote the models $\mathbf{N} = [1,1,2]\,, m = 3$ and $\mathbf{N} = [1,1,1]\,, m = 3$ by "nMrkv" and "Mrkv", respectively. From Mrkv to nMrkv, there is no change in the k_i's, although $\mathbf{A}^{\text{nMrkv}}$ is now 4×4:

$$\mathbf{A}^{\text{nMrkv}} = \begin{bmatrix} -k_1 & 0 & 0 & 0 \\ 0 & -k_2 & 0 & 0 \\ 0 & 0 & -k_3 & 1 \\ 0 & 0 & 0 & -k_3 \end{bmatrix}.$$

In addition, there are three more non-Markov parameters in $\mathbf{B}^{\text{nMrkv}}$:

$$\underbrace{\mathbf{B}^{\text{nMrkv}}}_{4\times 3} = \begin{bmatrix} \underbrace{\mathbf{B}^{\text{Mrkv}}}_{3\times 3} \\ \underbrace{\mathbf{B}_4^{\text{nMrkv}}}_{1\times 3} \end{bmatrix}.$$

Finally, there are three more parameters in λ_2^{nMrkv}, but there is no change in λ_1^{nMrkv}:

$$\underbrace{\lambda_2^{\text{nMrkv}}}_{3\times 4} = \left[\underbrace{\lambda_2^{\text{Mrkv}}}_{3\times 3}, \quad \underbrace{\lambda_{2,4}^{\text{nMrkv}}}_{3\times 1} \right],$$

$$\lambda_1^{\text{nMrkv}} = \lambda_1^{\text{Mrkv}}.$$

Therefore,

$$\mathbf{A}^{\mathbb{P},\text{nMrkv}} = \mathbf{A}^{\text{nMrkv}} - \mathbf{B}^{\text{nMrkv}} \lambda_2^{\text{nMrkv}},$$

$$\mu^{\text{nMrkv}} = \left(\mathbf{A}^{\mathbb{P},\text{nMrkv}} \right)^{-1} \mathbf{B}^{\text{nMrkv}} \lambda_1^{\text{Mrkv}}.$$

D Finding the best non-Markov model

Following the procedures outlined in the previous subsections, we estimate a three-factor Markov model along with the nineteen non-Markov models. We also estimate Joslin et al. (2013)'s VAR(4) model and Feunou and Fontaine (2016)'s VARMA(1,1) model, the details of which are presented in Appendix D. All models are estimated using only data from the in-sample periods. The AIC values of the different models are reported in Table 2. Among the three benchmark models (the three-factor Markov model, the VAR(4), and the VARMA(1,1)), the VAR(4) consistently has the lowest AIC values. However, we find that the AIC values of most of the non-Markov models to be even lower, especially those with five or six state variables. Generally speaking, non-Markov models with a larger number of states have lower AIC values. This indicates that the data exhibit a strong non-Markov property.[24]

[Insert Table 2 about here]

[24] Results using other information criteria, such as the SBIC and HQIC, are similar. These additional results are available upon request.

Given the parameter estimates from the in-sample periods, for the 20 continuous-time models (the Markov and non-Markov models specified under our framework), we filter the state variables from the observed zero yields in the full sample. For the two discrete-time models (VAR(4) and VARMA(1,1)), the state variables are the three principal components constructed using the weighting matrix estimated from the in-sample periods. Therefore, the first subsample of the state variables are in-sample values, while the second subsample are out-of-sample ones.

Following Cochrane and Piazzesi (2005), we focus on the average log excess holding period return across maturities between two to five years:

$$\begin{aligned}\overline{rx}_{t+1} &\equiv \frac{1}{4}\sum_{n=2}^{5} rx_{t+1}^{(n)},\\ rx_{t+1}^{(n)} &\equiv r_{t+1}^{(n)} - y_t^{(1)},\\ r_{t+1}^{(n)} &\equiv ny_t^{(n)} - (n-1)\, y_{t+1}^{(n-1)},\end{aligned}$$

where $y_t^{(n)}$ is the n-year zero yield at time t. Given the parameters and the state variables, the expected excess return $\mathbb{E}_t\left(\overline{rx}_{t+1}\right)$ is computed as

$$\begin{aligned}\mathbb{E}_t\left(\overline{rx}_{t+1}\right) = &\left(\frac{1}{4}\sum_{j=2}^{5} jy_t^{(j)} - y_t^{(1)}\right)\\ &-\frac{1}{4}\sum_{j=1}^{4}\left(j\varphi + \int_0^j \Theta^*(s)\,ds + \int_0^j \mathbf{C}(s)\,ds\left(\mathbb{1} - \exp\left(\mathbf{A}^{\mathbb{P}}\right)\right)\mu\right)\\ &-\frac{1}{4}\sum_{j=1}^{4}\int_0^j \mathbf{C}(s)\,ds\exp\left(\mathbf{A}^{\mathbb{P}}\right)Z_t,\end{aligned}$$

where we have used the relation between zero yields and state variables in (A5) and the conditional mean of Z_{t+1} given Z_t. Using the model-implied expected excess return $\mathbb{E}_t\left(\overline{rx}_{t+1}\right)$, we can calculate trading returns of the form $rn_{t+1} \equiv \overline{rx}_{t+1} \times \mathbb{E}_t\left(\overline{rx}_{t+1}\right)$.[25] We can then evaluate

[25] This is the return of the trading strategy considered in Cochrane and Piazzesi (2006). Specifically, we assume, for

the 22 models considered here based on their in- and out-of-sample adjusted returns (AdjRn= $\langle rn_t \rangle / \mathtt{std}(rn_t)$, i.e., the average of rn_t divided by its standard deviation) and cumulative returns (CumRn$_t$ $= \sum_{i=\Delta t}^{t/\Delta t} rn_i$, $\Delta t = 1/12$, i.e., the sum of rn_t over monthly intervals), as well as their in-sample R^2's from projecting the realized excess return onto its model-implied expectation.

We present the in-sample R^2's and the adjusted returns in Tables 3 and 4, respectively. Intuitively, a high R^2 means that the model can explain the excess bond returns well, and this should be a necessary condition for obtaining a mostly positive trading return rn_t and a high value for its risk-adjusted mean, which we denote as AdjRn. Indeed, this is confirmed upon a careful inspection of both tables. For example, Table 3 shows that for Germany (DE), the model $[1, 3, 2]$ has the highest in-sample R^2 among all models considered ($R^2 = 0.410$). In Table 4, $[1, 3, 2]$ also has the highest adjusted returns (AdjRn= 0.622) among all models for DE. The same inspection also reveals that there is no single model that outperforms for all countries. For example, in Table 3, VAR(4) has the highest in-sample R^2 for three countries (NZ, CH, and SE), while non-Markov models do so for the other six; in Table 4, the three-factor Markov model has the highest in-sample AdjRn for three countries (NZ, CH, and US), while VARMA(1,1) does so for one (AU) and non-Markov models round out the other five.

[Insert Tables 3 and 4 about here]

Figure 1 plots the cumulative returns (CumRns) for the in-sample periods. We include the three benchmark models (three-factor Markov, VAR(4), and VARMA(1,1)), and the best non-Markov specification according to its out-of-sample AdjRn (see Table 5). The graphs show that there is not a single GDTSM specification that consistently outperforms among the 22 continuous-time

simplicity, that we are able to trade a portfolio that gives a log annual return of $\overline{rx}_t$. We invest in $\mathbb{E}_t\left(\overline{rx}_{t+1}\right)$ units of this portfolio every month and close the position after holding it for 12 months. We long the portfolio if $\mathbb{E}_t\left(\overline{rx}_{t+1}\right) > 0$ and short the portfolio if $\mathbb{E}_t\left(\overline{rx}_{t+1}\right) < 0$. The return of this "trading rule" is the product of the excess return and its expected value one year prior:

$$rn_t \equiv \overline{rx}_t \times \mathbb{E}_{t-1}\left(\overline{rx}_t\right).$$

and discrete-time GDTSMs. The best performing model seems to be spread evenly across the countries among VAR(4), VARMA(1,1), the three-factor Markov model, and the best three-factor non-Markov specification. Therefore, while the information criteria clearly show an advantage in using non-Markov models to describe bond yields, this advantage is less obvious based on the trading profits from a strategy that exploits the predictability of bond excess returns.

[Insert Figure 1 about here]

In model selection, however, one must be cautious about relying too much on in-sample fitting. Ultimately, whether a model successfully captures the true data generating process can only be tested with out-of-sample model predictions. It is in this context that problems with overfitting are easily revealed. Therefore, we also examine the adjusted returns and cumulative returns of the trading strategy in an out-of-sample context, keeping model parameters fixed at their in-sample estimates while continuing to filter the state variables and predict the bond excess returns. Since our out-of-sample periods are five, nine, or 15 years long, depending on the country, we subject the models to a rather severe form of out-of-sample testing. This is intentional, however, since our objective is to see which model is capable of capturing the true term structure dynamics in the long-run.[26]

We present the out-of-sample AdjRns in Table 5, in which we highlight, for each country, one of the 22 GDTSMs that yields the highest AdjRn. In two of the countries this is the three-factor Markov specification, while one of our non-Markov specifications is the best for the other seven countries. Comparing the in-sample AdjRns from Table 4 with the out-of-sample AdjRns from Table 5, we find that the performance of the three benchmark models declines significantly from

[26] The common practice for out-of-sample testing is to do re-calibration using a rolling window of historical observations. However, the re-calibration of model parameters tends to mask the differences among the models. Our specifications are also more time-consuming to estimate compared to, say, the Joslin et al. (2013) model, and it would be infeasible for us to conduct rolling window estimation across the 19 non-Markov models and nine countries.

in-sample to out-of-sample. Since we do not consider transaction costs when computing the trading returns, we impose a somewhat arbitrary risk-adjusted return of 30 percent or more for it to be qualified as economically significant (other cutoffs yield similar insights). Consequently, while VAR(4), VARMA(1,1), and the three-factor Markov model all yield economically significant in-sample AdjRns for all nine countries, VAR(4) no longer does so for any country out-of-sample, VARMA(1,1) does so for only one (US), and the three-factor Markov model, two (NZ and SE).

[Insert Table 5 about here]

In contrast, the situation with non-Markov models is more encouraging. Take the model $[4, 1, 1]$ as an example. It generates economically significant in-sample AdjRns for six countries (AU, JP, DE, SE, CA, and US). Out of these six cases, five (AU, DE, SE, CA, and US) continue to have economically significant out-of-sample AdjRns. While the performance of the other non-Markov specifications is slightly weaker, even a randomly chosen non-Markov specification among the 19 seems to perform better out-of-sample than the three benchmark models. One way to see the superior out-of-sample performance of the non-Markov models as a group is to count the number of cases in which they generate economically significant AdjRns. This is equal to 3 (AU), 7 (JP), 0 (NZ), 14 (CH), 14 (DE), 5 (SE), 9 (CA), 8 (GB), and 13 (US). Hence, with the exception of NZ, we have a large number of non-Markov specifications to choose from that can potentially capture the underlying term structure dynamics.

To be sure that the best-performing non-Markov specification is not selected by chance alone because we have included a large number of such specifications, we also measure the average performance of the models. Specifically, we compute for each model the fraction of cases with positive in-sample AdjRns that also have out-of-sample AdjRns larger than 30 percent. Across all non-Markov models, this ratio averages 46 percent. In contrast, the average across the three benchmark models is only 11 percent. This confirms that even a randomly selected non-Markov specification among the 19 that we have considered would outperform the three benchmark models.

The advantage of our non-Markov setup is also confirmed by the out-of-sample CumRn plots in Figure 2, which look distinctly different from their in-sample counterparts in Figure 1. For the majority of the countries, the non-Markov specification selected to yield the best out-of-sample AdjRn also yields the highest CumRn. In summary, although our non-Markov models perform similarly in-sample as the three benchmark models, their superior out-of-sample performance indicates that the non-Markov feature captures a crucial aspect of term structure dynamics.

[Insert Figure 2 about here]

E More states than factors: A potential feature embedded in fundamentals

As Table 5 shows, a three-factor non-Markov model with six states delivers the best out-of-sample performance in four countries (CH, DE, CA, and US), a model with five states does so in one country (GB), and a model with four states does so in two countries (AU and JP). This suggests that our flexible and parsimonious framework for specifying GDTSMs is indeed helpful for capturing the non-Markov property across many different bond markets.

Under a general equilibrium setting, the dynamics of interest rates is explicitly determined by the dynamics of the fundamental economic variables, such as investment and production (e.g., Cox et al., 1985; Longstaff and Schwartz, 1992). Indeed, Jin and Glasserman (2001) show that every HJM model can arise as the equilibrium term structure in a Cox-Ingersoll-Ross production economy. Therefore, the empirical fact that deterministically varying states in the dynamics of interest rates can significantly forecast excess bond returns indicates that a similar non-Markov structure might also be present in the economic fundamentals. One way in which this can occur has been suggested by Cox et al. (1981) (footnote 34): "One possible justification (for why past interest rates are plausible state variables in a rational expectations equilibrium) arises when investment is not readily reversible so that past interest rates are still reflected in the current production function. Changes in the interest rate will then be affected by past rates as these investments disappear or are abandoned." Merton (1973) also mentions that the Markov property of "the stochastic processes describing the opportunity set and its changes" is rather general in the sense that the stochastic processes describing returns can be non-Markov, but by including supplementary variables, the entire (expanded) set is once again Markov. The extra states in our models echo the very idea of "supplementary variables." However, to the best of our knowledge, there are only a few studies, mostly theoretical, in this direction.[27]

V Conclusion

Motivated by the strong evidence of non-Markov state variables in the term structure of interest rates, as well as the existing literature on bond yields spanning lagged (macro) risk factors, we develop a systematic approach for building non-Markov GDTSMs. This approach not only inherits the canonical features emphasized in the recent literature, e.g., Joslin et al. (2011), but also offers great flexibility in specifying non-Markov dynamics for the state variables under both $\mathbb{Q}$ and $\mathbb{P}$. Exploiting the flexibility of our approach, we conduct a specification analysis to examine the ability of our non-Markov GDTSMs to forecast bond excess returns out-of-sample. We find that in a majority of the bond markets (seven out of nine, including the U.S.), the traditional three-factor Markov model cannot produce economically significant trading profits. In contrast, in most of the markets (eight out of nine), there are at least three non-Markov specifications producing economically significant trading profits, with different specifications producing the best results in different bond markets. In five of the nine markets, at least two or three deterministically varying state vari-

[27] One such example is Dumas et al. (2009). In their model, four state variables are driven by two independent random sources.

ables are needed to capture the non-Markov property in a model with three independent random sources (factors). This suggests that the non-Markov property can be strong and non-trivial to model. Collectively, the empirical evidence presented in our paper suggests that the exploration of non-Markov properties within a general equilibrium framework can be a fruitful avenue for future research.

Table 1: Data summary

This table summarizes the start and end dates of the data. The dates are of the "yyyymm" format. The lengths of the data in terms of the number of months are also reported for the in-sample and the full sample periods.

	Asia Pacific			Continental Europe			NA and UK Group		
	AU	JP	NZ	CH	DE	SE	CA	GB	US
Start	198702	198501	199001	198801	197301	199212	198601	197901	197111
In-sample End	200004	200004	200404	200004	199404	200404	200004	199404	199404
Full-sample End	200904	200904	200904	200904	200904	200904	200904	200904	200904
In-sample mths	159	184	172	148	256	137	172	184	270
Full-sample mths	268	293	233	257	437	198	281	365	451

Table 2: In-sample AIC

This table reports the in-sample AIC of the Markov and non-Markov models. The number of states and factors is indicated in the first two columns. The model specification is indicated in the third column.

			Asia Pacific			Continental Europe			NA and UK Group		
			AU	JP	NZ	CH	DE	SE	CA	GB	US
Markov	3-factor	[1 1 1]	-2.636	-3.101	-2.918	-2.481	-3.914	-2.303	-2.642	-2.729	-4.380
non-Markov (3-factor)	4-state	[1 1 2]	-2.639	-3.021	-2.969	-2.524	-4.176	-2.314	-2.459	-2.582	-4.585
		[1 2 1]	-2.777	-3.100	-2.988	-2.565	-4.055	-2.243	-2.652	-2.621	-4.665
		[2 1 1]	-2.843	-3.191	-2.838	-2.526	-4.118	-2.263	-2.588	-2.724	-4.698
	5-state	[1 1 3]	-2.835	-3.399	-3.098	-2.492	-3.929	-2.254	-2.575	-2.743	-4.736
		[1 2 2]	-2.831	-3.496	-3.070	-2.593	-4.313	-2.301	-3.012	-2.891	-5.093
		[1 3 1]	-3.010	-3.430	-3.009	-2.827	-4.466	-2.145	-3.124	-3.011	-4.880
		[2 1 2]	-3.016	-3.296	-3.021	-2.593	-4.613	-2.400	-2.795	-2.963	-5.137
		[2 2 1]	-2.997	-3.064	-3.089	-2.834	-4.724	-2.342	-2.857	-2.882	-5.176
		[3 1 1]	-2.979	-3.454	-3.130	-2.678	-4.522	-2.521	-2.887	-2.950	-5.029
	6-state	[1 1 4]	-3.290	-3.675	-3.152	-2.821	-4.215	-2.537	-2.741	-3.267	-4.864
		[1 2 3]	-3.050	-3.735	-2.856	-2.603	-4.435	-2.131	-3.149	-3.031	-5.338
		[1 3 2]	-3.137	-3.743	-3.030	-2.911	-4.787	-2.060	-3.355	-3.193	-5.345
		[1 4 1]	-3.380	-3.765	-3.183	-3.123	-4.529	-2.176	-3.268	-3.391	-5.605
		[2 1 3]	-3.060	-3.477	-3.174	-2.808	-4.364	-2.327	-3.299	-2.983	-5.166
		[2 2 2]	-3.162	-3.598	-3.152	-3.138	-4.770	-2.598	-3.150	-3.175	-5.439
		[2 3 1]	-3.185	-3.578	-3.341	-2.750	-4.900	-2.509	-3.165	-3.418	-5.686
		[3 1 2]	-3.171	-3.712	-3.407	-3.016	-4.606	-2.527	-2.964	-3.141	-5.358
		[3 2 1]	-3.140	-3.450	-3.311	-3.054	-5.071	-2.568	-3.113	-3.252	-5.396
		[4 1 1]	-3.396	-3.758	-4.106	-3.295	-4.545	-2.357	-3.208	-3.322	-5.739
Discrete-time non-Markov	3-factor	VAR(4)	-2.750	-3.228	-3.087	-2.579	-4.005	-2.409	-2.693	-2.826	-4.615
		VARMA(1,1)	-2.259	-2.663	-2.453	-2.154	-3.419	-1.971	-2.385	-2.473	-3.822

Table 3: In-sample R-squared's

This table reports the in-sample R^2's of the Markov and non-Markov models for the nine bond markets. The number of states and factors is indicated in the first two columns. The model specification is indicated in the third column. The numbers reported represent the R^2's of projecting the realized excess return onto the model-implied expected excess return.

			Asia Pacific			Continental Europe			NA and UK Group		
			AU	JP	NZ	CH	DE	SE	CA	GB	US
Markov	3-factor	[1 1 1]	0.087	0.188	0.297	0.594	0.175	0.116	0.203	0.179	0.304
non-Markov (3-factor)	4-state	[1 1 2]	0.011	0.032	0.090	0.383	0.099	0.066	0.151	0.108	0.151
		[1 2 1]	0.041	0.147	0.002	0.193	0.095	0.097	0.106	0.062	0.119
		[2 1 1]	0.007	0.225	0.223	0.095	0.239	0.274	0.133	0.116	0.208
	5-state	[1 1 3]	0.009	0.111	0.058	0.389	0.106	0.206	0.127	0.417	0.197
		[1 2 2]	0.057	0.253	0.001	0.171	0.097	0.035	0.130	0.005	0.226
		[1 3 1]	0.039	0.089	0.177	0.107	0.009	0.127	0.137	0.121	0.246
		[2 1 2]	0.026	0.069	0.211	0.019	0.095	0.028	0.191	0.240	0.073
		[2 2 1]	0.038	0.282	0.024	0.017	0.046	0.050	0.066	0.031	0.216
		[3 1 1]	0.009	0.062	0.276	0.040	0.008	0.000	0.092	0.044	0.092
	6-state	[1 1 4]	0.092	0.012	0.408	0.287	0.143	0.012	0.127	0.004	0.227
		[1 2 3]	0.028	0.162	0.155	0.175	0.159	0.051	0.114	0.040	0.121
		[1 3 2]	0.029	0.120	0.239	0.144	0.410	0.121	0.103	0.061	0.091
		[1 4 1]	0.004	0.084	0.176	0.225	0.000	0.021	0.032	0.001	0.241
		[2 1 3]	0.010	0.147	0.341	0.086	0.167	0.001	0.035	0.430	0.092
		[2 2 2]	0.108	0.167	0.009	0.001	0.068	0.003	0.060	0.027	0.204
		[2 3 1]	0.011	0.142	0.013	0.013	0.026	0.005	0.079	0.065	0.325
		[3 1 2]	0.000	0.046	0.043	0.299	0.325	0.194	0.222	0.067	0.232
		[3 2 1]	0.178	0.021	0.352	0.023	0.003	0.043	0.050	0.003	0.224
		[4 1 1]	0.021	0.251	0.056	0.424	0.072	0.323	0.150	0.025	0.181
Discrete time non-Markov	3-factor	VAR(4)	0.064	0.207	0.505	0.649	0.242	0.359	0.155	0.207	0.287
		VARMA(1,1)	0.098	0.058	0.453	0.347	0.226	0.024	0.078	0.178	0.281

Table 4: In-sample AdjRns

This table reports the in-sample *AdjRn*s of the Markov and non-Markov models for the nine bond markets. The number of states and factors is indicated in the first two columns. The model specification is indicated in the third column. The results presented are based on the model-implied expected excess return.

			Asia Pacific			Continental Europe			NA and UK Group		
			AU	JP	NZ	CH	DE	SE	CA	GB	US
Markov	3-factor	[1 1 1]	0.476	0.670	0.464	0.815	0.438	0.423	0.516	0.303	0.433
non-Markov (3-factor)	4-state	[1 1 2]	0.134	0.534	0.350	0.733	0.372	0.313	0.473	0.258	0.370
		[1 2 1]	0.257	0.050	0.010	0.491	0.400	0.213	0.412	0.027	0.322
		[2 1 1]	0.287	0.681	0.138	-0.247	0.476	0.469	0.459	-0.311	0.383
	5-state	[1 1 3]	0.379	0.095	0.285	0.659	0.379	0.153	0.420	0.404	0.408
		[1 2 2]	0.415	0.436	-0.112	0.461	0.370	-0.283	0.436	-0.069	0.278
		[1 3 1]	0.355	-0.290	0.393	0.352	0.270	0.214	0.451	0.227	0.285
		[2 1 2]	0.102	0.648	0.191	-0.104	-0.378	-0.291	0.452	0.333	0.283
		[2 2 1]	0.358	-0.349	-0.368	0.141	0.354	0.227	0.337	-0.224	0.313
		[3 1 1]	0.369	-0.540	-0.226	0.217	0.338	0.613	0.377	0.240	-0.251
	6-state	[1 1 4]	0.366	0.695	0.272	0.537	0.399	0.290	0.399	-0.043	0.359
		[1 2 3]	0.278	0.267	0.386	0.470	0.451	0.163	0.429	-0.242	0.256
		[1 3 2]	0.391	0.150	0.372	0.375	0.622	0.265	0.390	0.175	0.251
		[1 4 1]	0.255	-0.043	0.049	0.527	0.216	-0.311	0.357	-0.129	0.280
		[2 1 3]	0.288	0.788	0.444	-0.324	-0.301	0.396	0.300	0.468	0.274
		[2 2 2]	0.050	0.500	-0.176	-0.006	0.384	0.278	0.121	0.202	0.324
		[2 3 1]	-0.050	0.280	-0.276	-0.071	0.351	0.379	0.143	-0.110	0.388
		[3 1 2]	0.213	0.552	-0.252	-0.177	0.593	0.647	0.544	0.213	0.309
		[3 2 1]	-0.179	-0.600	-0.032	-0.165	0.315	0.176	0.385	0.135	0.329
		[4 1 1]	0.391	0.337	-0.151	-0.665	0.399	0.629	0.460	0.124	0.314
Discrete time non-Markov	3-factor	VAR(4)	0.385	0.488	0.386	0.747	0.444	0.595	0.385	0.331	0.413
		VARMA(1,1)	0.481	0.477	0.458	0.377	0.459	0.381	0.402	0.309	0.379

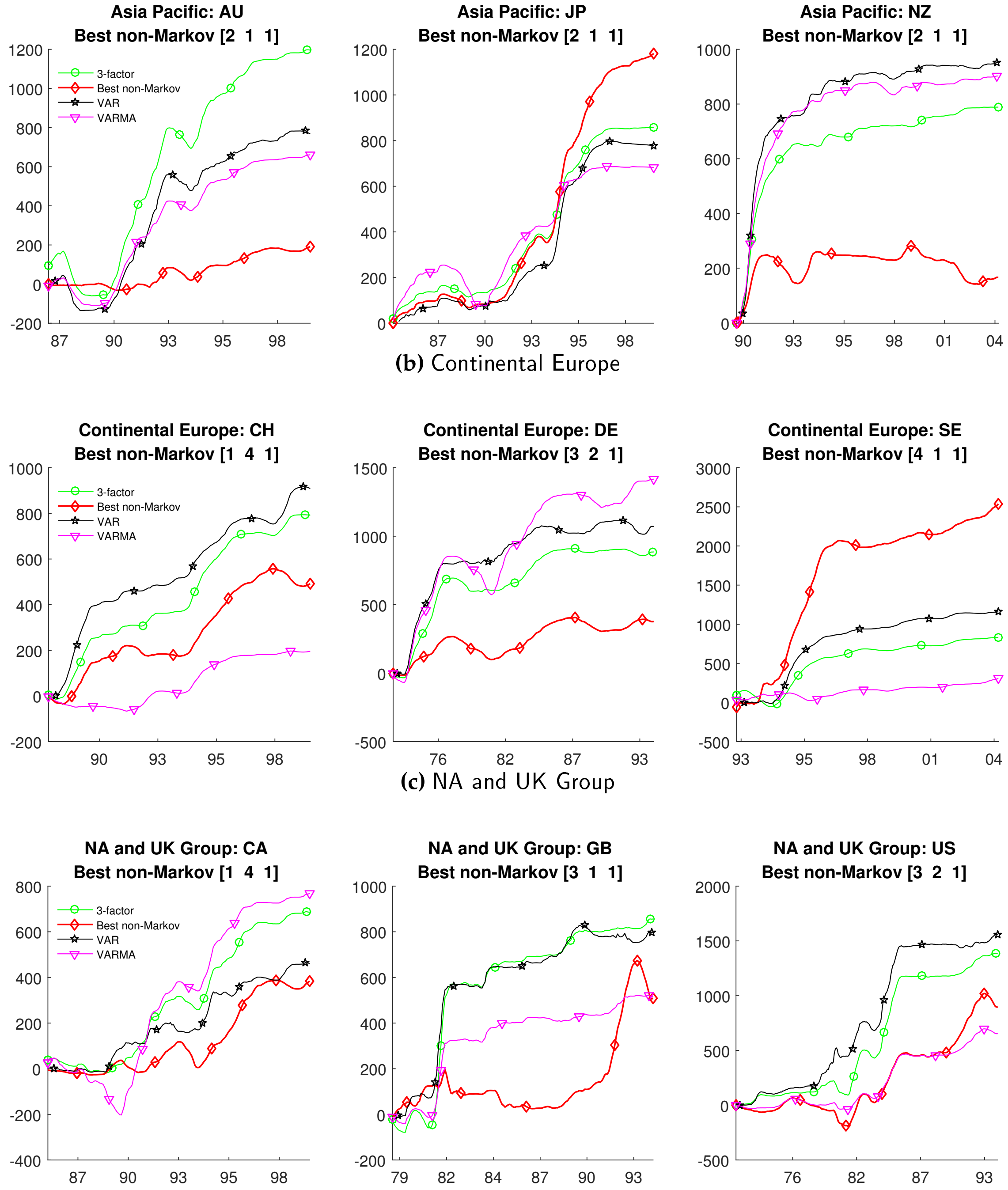
Table 5: Out-of-sample AdjRns

This table reports the out-of-sample *AdjRn*s of the Markov and non-Markov models for the nine bond markets. The number of states and factors is indicated in the first two columns. The model specification is indicated in the third column. The results presented are based on the model-implied expected excess return. The best result in each column is highlighted in gray.

			Asia Pacific			Continental Europe			NA and UK Group		
			AU	JP	NZ	CH	DE	SE	CA	GB	US
Markov	3-factor	[1 1 1]	-0.004	-0.438	0.633	-0.260	0.196	0.469	-0.338	-0.120	-0.330
non-Markov (3-factor)	4-state	[1 1 2]	0.274	0.802	0.016	0.133	0.386	-0.256	0.299	-0.247	-0.137
		[1 2 1]	-0.101	-0.781	0.076	0.591	0.540	-0.325	0.462	-0.363	0.364
		[2 1 1]	0.377	0.877	0.293	0.429	0.440	-0.170	0.447	0.394	0.470
	5-state	[1 1 3]	0.036	-0.776	0.107	0.471	0.634	-0.346	0.159	-0.252	-0.222
		[1 2 2]	0.328	-0.666	0.040	0.436	0.146	-0.336	0.301	0.443	0.565
		[1 3 1]	-0.052	-0.797	0.096	0.555	0.657	-0.274	0.622	0.135	0.303
		[2 1 2]	0.238	0.839	0.037	-0.523	-0.125	-0.365	-0.045	-0.314	-0.135
		[2 2 1]	0.096	0.721	0.018	0.567	0.635	0.436	-0.463	0.428	-0.348
		[3 1 1]	0.265	-0.812	0.029	0.464	0.664	0.406	0.366	0.524	0.464
	6-state	[1 1 4]	-0.183	0.816	0.065	0.622	0.546	-0.376	0.074	0.201	0.144
		[1 2 3]	-0.141	-0.755	0.069	0.405	0.462	0.081	0.202	-0.448	0.499
		[1 3 2]	0.270	-0.783	0.076	0.572	-0.118	-0.448	0.188	0.068	0.529
		[1 4 1]	-0.228	-0.793	0.008	0.634	0.625	-0.351	0.994	0.425	0.329
		[2 1 3]	0.215	0.555	0.073	0.459	0.298	0.430	0.904	-0.181	-0.374
		[2 2 2]	-0.167	-0.678	-0.006	-0.558	0.528	-0.100	-0.398	-0.458	0.547
		[2 3 1]	-0.215	-0.762	0.011	-0.537	0.582	0.410	-0.766	0.394	0.578
		[3 1 2]	-0.179	0.763	0.070	-0.700	-0.449	0.209	0.003	-0.176	0.534
		[3 2 1]	-0.234	-0.788	0.033	0.525	0.737	0.253	0.950	0.479	0.596
		[4 1 1]	0.349	-0.189	0.033	0.373	0.610	0.452	0.709	0.519	0.501
Discrete time non-Markov	3-factor	VAR(4)	-0.124	-0.538	0.296	-0.514	-0.272	-0.028	-0.506	-0.069	-0.378
		VARMA(1,1)	0.248	-0.418	0.126	0.284	0.140	0.228	-0.493	0.206	0.424

Figure 1: In-sample CumRns: Best non-Markov model vs. benchmarks

The plots in this figure present the in-sample *CumRn*s over time for the best non-Markov model (among the 19 specifications), the benchmark models (three-factor Markov model, VAR(4), and VARMA(1,1)) for the nine bond markets. Panels (a), (b), and (c) present results of the Asia Pacific, Continental Europe, and NA and UK Group, respectively. The *CumRn*s are in basis points. All the x-axes are year in the format of "yy".



(a) Asia Pacific

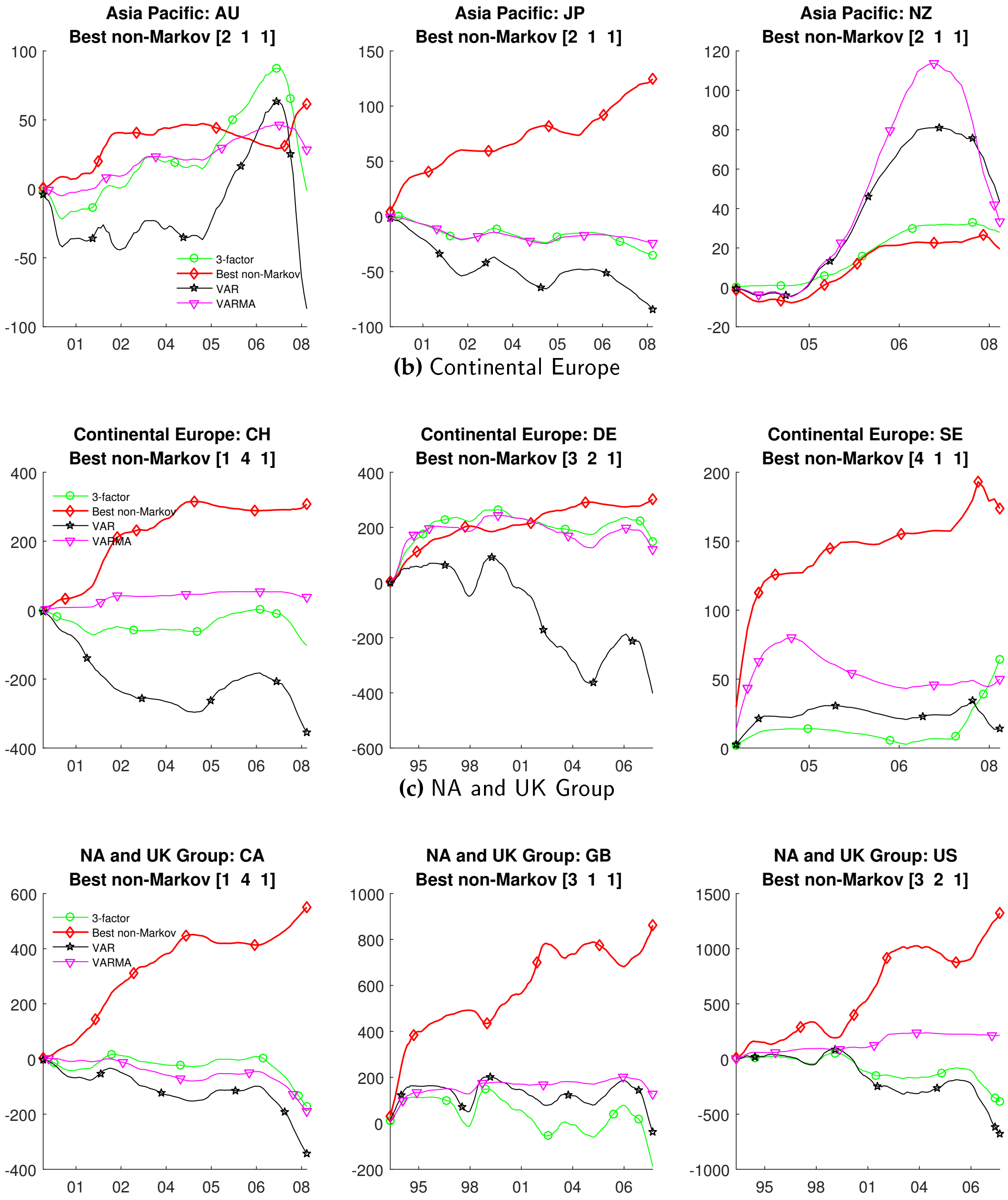
(b) Continental Europe

(c) NA and UK Group

Figure 2: Out-of-sample CumRns: Best non-Markov model vs. benchmarks

The plots in this figure present the out-of-sample *CumRn*s over time for the best non-Markov model (among the 19 specifications), the benchmark models (three-factor Markov models, VAR(4), and VARMA(1,1)). Panels (a), (b), and (c) present results of the Asia Pacific, Continental Europe, and NA and UK Group, respectively. The *CumRn*s are in basis points. All the x-axes are year in the format of "yy".

(a) Asia Pacific



(b) Continental Europe

(c) NA and UK Group

Appendices

A Time-homogeneous forward curves

By taking the limit as $t \to \infty$ in (2), the forward rate can be rewritten as $r(t,x) = \varphi + \Theta^*(x) + \mathbf{C}_0 \exp(\mathbf{A}x) Z_t$, where

$$\Theta^*(x) \equiv \lim_{t\to\infty} \Theta(t,x) \tag{A1}$$

$$= \mathbf{C}(x)\left(\mathbf{A}^{-1}\mathbf{B}\mathbf{B}^\intercal(\mathbf{A}^\intercal)^{-1}\right)\mathbf{C}_0^\intercal - \frac{1}{2}\mathbf{C}(x)\left(\mathbf{A}^{-1}\mathbf{B}\mathbf{B}^\intercal(\mathbf{A}^\intercal)^{-1}\right)\mathbf{C}(x)^\intercal, \tag{A2}$$

$$\varphi \equiv \lim_{t\to\infty} r(0,t+x). \tag{A3}$$

In Appendix B, we show that the second equality holds for any invertible $\mathbf{A}$. The zero-coupon bond price is given by:

$$P(t,t+x) = \exp\left(-\int_0^x r(t,s)\,ds\right) = \exp\left(\mathbf{H}(x) - \mathbf{F}(x)^\intercal Z_t\right), \tag{A4}$$

where

$$\mathbf{H}(x) = -\varphi x - \int_0^x \Theta^*(s)\,ds,$$

$$\mathbf{F}(x)^\intercal = \mathbf{C}_0\int_0^x \exp(\mathbf{A}s)\,ds = (\mathbf{C}(x) - \mathbf{C}_0)\,\mathbf{A}^{-1}.$$

These results can also be derived using the traditional GDTSM approach by starting from the short rate specification. That is, the short rate is:

$$r(t,0) = \varphi + \Theta^*(0) + \mathbf{C}_0 Z_t, \text{ where } dZ_t = \mathbf{A}Z_t dt + \mathbf{B}dW_t, \quad Z_0 = \mathbf{0}.$$

Then, $\mathbf{H}(x)$ and $\mathbf{F}(x)$ can be solved from the following ordinary differential equations:

$$\frac{d\mathbf{H}(x)}{dx} = \frac{1}{2}\mathbf{F}(x)^{\intercal}\mathbf{B}\mathbf{B}^{\intercal}\mathbf{F}(x) - (\varphi + \Theta^{*}(0)), \quad \frac{d\mathbf{F}(x)}{dx} = \mathbf{A}^{\intercal}\mathbf{F}(x) + \mathbf{C}_0^{\intercal},$$

with the boundary conditions $\mathbf{H}(0) = 0$, and $\mathbf{F}(0) = \mathbf{0}_{m\times 1}$.

Therefore, following (A4), the model-implied n-year zero yield at time t can be expressed as:

$$y_t^{(n)} = \varphi + \frac{\int_0^n \Theta^{*}(s)\,ds}{n} + \frac{\int_0^n \mathbf{C}(s)\,ds}{n} Z_t. \tag{A5}$$

B Derivation of $\Theta^{*}(x)$

By $\Theta(t,x) = \int_0^t \sigma(x+t-s)\int_0^{x+t-s}\sigma(\tau)^{\intercal}\,d\tau ds$ and $\sigma(x) = \mathbf{C}(x)\mathbf{B}$, we have:

$$\begin{aligned}
\Theta(t,x) &= \mathbf{C}(x)\left[\int_0^t \exp(\mathbf{A}(t-s))\,\mathbf{B}\mathbf{B}^{\intercal}\int_0^{x+t-s}\exp(\mathbf{A}^{\intercal}\tau)\,d\tau ds\right]\mathbf{C}_0^{\intercal} \\
&= \mathbf{C}(x)\int_0^t \exp(\mathbf{A}(t-s))\,\mathbf{B}\mathbf{B}^{\intercal}\begin{bmatrix} \int_0^x \exp(\mathbf{A}^{\intercal}\tau)\,d\tau + \\ \\ \int_0^{t-s}\exp(\mathbf{A}^{\intercal}\tau)\,d\tau\exp(\mathbf{A}^{\intercal}x)\end{bmatrix} ds\mathbf{C}_0^{\intercal} \\
&= \mathbf{C}(x)\int_0^t \exp(\mathbf{A}(t-s))\,ds\mathbf{B}\mathbf{B}^{\intercal}\int_0^x \exp(\mathbf{A}^{\intercal}\tau)\,d\tau\mathbf{C}_0^{\intercal} \\
&+ \mathbf{C}(x)\int_0^t \exp(\mathbf{A}(t-s))\,\mathbf{B}\mathbf{B}^{\intercal}\int_0^{t-s}\exp(\mathbf{A}^{\intercal}\tau)\,d\tau ds\mathbf{C}(x)^{\intercal} \\
&= \mathbf{C}(x)\int_0^t \exp(\mathbf{A}(t-s))\,\mathbf{B}\mathbf{B}^{\intercal}\exp(\mathbf{A}^{\intercal}(t-s))\,ds\,(\mathbf{A}^{\intercal})^{-1}\mathbf{C}(x)^{\intercal} \\
&- \mathbf{C}(x)\int_0^t \exp(\mathbf{A}(t-s))\,\mathbf{B}\mathbf{B}^{\intercal}ds\,(\mathbf{A}^{\intercal})^{-1}\mathbf{C}_0^{\intercal}.
\end{aligned}$$

Then,

$$\begin{aligned}\Theta^*(x) &= \mathbf{C}(x)\left[\lim_{t\to+\infty}\int_0^t \exp\left(\mathbf{A}(t-s)\right)\mathbf{B}\mathbf{B}^\intercal \exp\left(\mathbf{A}^\intercal(t-s)\right)ds\right]\left(\mathbf{A}^\intercal\right)^{-1}\mathbf{C}(x)^\intercal \\ &- \mathbf{C}(x)\left[\lim_{t\to+\infty}\int_0^t \exp\left(\mathbf{A}(t-s)\right)ds\right]\mathbf{B}\mathbf{B}^\intercal\left(\mathbf{A}^\intercal\right)^{-1}\mathbf{C}_0^\intercal \\ &= \mathbf{C}(x)\left(\mathbf{A}^{-1}\mathbf{B}\mathbf{B}^\intercal\left(\mathbf{A}^\intercal\right)^{-1}\right)\mathbf{C}_0^\intercal - \frac{1}{2}\mathbf{C}(x)\left(\mathbf{A}^{-1}\mathbf{B}\mathbf{B}^\intercal\left(\mathbf{A}^\intercal\right)^{-1}\right)\mathbf{C}(x)^\intercal.\end{aligned}$$

The second equation above is by the fact that $\mathrm{Y} \equiv \int_0^\infty \exp\left(\mathbf{A}(t-s)\right)\mathbf{B}\mathbf{B}^\intercal \exp\left(\mathbf{A}^\intercal(t-s)\right)ds$ satisfies the Lyapunov equation:

$$\mathbf{A}\mathrm{Y} + \mathrm{Y}\mathbf{A}^\intercal = -\mathbf{B}\mathbf{B}^\intercal.$$

Therefore,

$$\mathbf{C}(x)\,\mathrm{Y}\left(\mathbf{A}^\intercal\right)^{-1}\mathbf{C}(x)^\intercal + \mathbf{C}(x)\,\mathbf{A}^{-1}\mathrm{Y}\mathbf{C}(x)^\intercal = -\mathbf{C}(x)\left(\mathbf{A}^{-1}\mathbf{B}\mathbf{B}^\intercal\left(\mathbf{A}^\intercal\right)^{-1}\right)\mathbf{C}(x)^\intercal,$$

which means that

$$\mathbf{C}(x)\,\mathrm{Y}\left(\mathbf{A}^\intercal\right)^{-1}\mathbf{C}(x)^\intercal = -\frac{1}{2}\mathbf{C}(x)\left(\mathbf{A}^{-1}\mathbf{B}\mathbf{B}^\intercal\left(\mathbf{A}^\intercal\right)^{-1}\right)\mathbf{C}(x)^\intercal,$$

since $\mathbf{C}(x)\,\mathrm{Y}\left(\mathbf{A}^\intercal\right)^{-1}\mathbf{C}(x)^\intercal$ and $\mathbf{C}(x)\,\mathbf{A}^{-1}\mathrm{Y}\mathbf{C}(x)^\intercal$ are both scalars, and are transposes of each other.

C Proofs

A Proof of Proposition 1

It is trivial that any FDR with $n > m$ can be transformed into one with $\mathbf{B}$ being

$$\begin{bmatrix} \underbrace{B_1}_{m\times m} \\ \\ \underbrace{\mathbf{0}}_{(n-m)\times m} \end{bmatrix}.$$

So, without loss of generality, we consider the following Z_t as representative of any FDR with $n > m$:

$$dZ_t = d\begin{bmatrix} \underbrace{Z_{1,t}}_{m\times 1} \\ \\ \underbrace{Z_{2,t}}_{(n-m)\times 1} \end{bmatrix} = \begin{bmatrix} \underbrace{A_{11}}_{m\times m} & \underbrace{A_{12}}_{m\times(n-m)} \\ \\ \underbrace{A_{21}}_{(n-m)\times m} & \underbrace{A_{22}}_{(n-m)\times(n-m)} \end{bmatrix} \begin{bmatrix} \underbrace{Z_{1,t}}_{m\times 1} \\ \\ \underbrace{Z_{2,t}}_{(n-m)\times 1} \end{bmatrix} dt + \begin{bmatrix} \underbrace{B_1}_{m\times m} \\ \\ \underbrace{\mathbf{0}}_{(n-m)\times m} \end{bmatrix} \underbrace{dW_t}_{m\times 1}.$$

Therefore, the dynamics of $Z_{2,t}$ is given by an ODE:

$$dZ_{2,t} = (A_{21}Z_{1,t} + A_{22}Z_{2,t})\,dt,\ Z_{2,0} = 0. \tag{A6}$$

That is, $Z_{2,t} = \int_0^t A_{21}Z_{1,s}ds + \int_0^t A_{22}Z_{2,s}ds$. Given this, the dynamic of $Z_{1,t}$ is:

$$dZ_{1,t} = \left(A_{11}Z_{1,t} + A_{12}A_{21}\int_0^t Z_{1,s}ds + A_{12}A_{22}\int_0^t Z_{2,s}ds\right)dt + B_1 dW_t.$$

Apparently, the drift term of $Z_{1,t}$ contains integrals of its own and $Z_{2,t}$'s historical values.

We now show that $Z_{2,t}$ is an exponentially-weighted average of $Z_{1,t}$. We guess, and later verify, that $Z_{2,t}$ takes the form

$$Z_{2,t} = e^{A_{22}t}\phi\left(t\right).$$

Therefore, we have

$$dZ_{2,t} = A_{22}e^{A_{22}t}\phi\left(t\right)dt + e^{A_{22}t}d\phi\left(t\right). \tag{A7}$$

Comparing (A7) with (A6), it is clear that $\phi\left(t\right)$ solves the following ODE:

$$d\phi\left(t\right) = e^{-A_{22}t}A_{21}Z_{1,t}dt.$$

Thus $\phi\left(t\right) = \int_0^t e^{-A_{22}s}A_{21}Z_{1,s}ds$ and $Z_{2,t} = \int_0^t e^{A_{22}(t-s)}A_{21}Z_{1,s}ds$. Therefore, the drift of $Z_{1,t}$ generally depends on its own history, showing that the first m states are non-Markov on their own.

B Proof of Theorem 1

To show that the triplet $\{\mathbf{A}, \mathbf{B}, \mathbf{C}\left(x\right)\}$ is a realization, according to Björk and Gombani, 1999, Proposition 3.1, we only need to demonstrate that:

$$\mathbf{C}\left(0\right)\exp\left(\mathbf{A}x\right) = \mathbf{C}\left(x\right) = \left[\left[1, x, \cdots, x^{n_i-1}\right]e^{-k_i x}\right]_{i=1}^{I}.$$

First, we rewrite $\left[\left[1, x, \cdots, x^{n_i-1}\right]e^{-k_i x}\right]_{i=1}^{I}$ in matrix form:

$$\left[\left[1, x, \cdots, x^{n_i-1}\right]e^{-k_i x}\right]_{i=1}^{I} = \text{Poly}\left(\mathbf{N}, x\right)\text{ExpM}\left(\mathbf{K}, \mathbf{N}, x\right),$$

where $\mathbf{N} = [n_1, n_2, \cdots, n_I]$ is a $1 \times I$ row vector with elements of natural numbers, $\mathbf{K} = [k_1, k_2, \cdots, k_I]$ is a $1 \times I$ row vector with elements of positive real numbers, $\text{ExpM}\left(\mathbf{K}, \mathbf{N}, x\right) : \left(\boldsymbol{R}^{1\times I}, \boldsymbol{N}^{1\times I}, \boldsymbol{R}_+\right) \to$

$\boldsymbol{R}_+^{n\times n}$, $n = \sum_{i=1}^{I} n_i$, is a matrix function defined as:

$$\mathrm{ExpM}\left(\mathbf{K}, \mathbf{N}, x\right) \equiv \begin{bmatrix} ExpM_1 & & \cdots & 0 \\ & ExpM_2 & & \vdots \\ \vdots & & \ddots & \\ 0 & \cdots & & ExpM_I \end{bmatrix},$$

$$ExpM_i \equiv \exp\left\{-\underbrace{diag\left[k_i, \cdots, k_i\right]}_{n_i\times n_i} x\right\},$$

where $diag[\cdots]$ is a compact notation for a diagonal matrix, and $\mathrm{Poly}\left(\mathbf{N}, x\right) : \left(\boldsymbol{N}^{1\times I}, \boldsymbol{R}_+\right) \to \boldsymbol{R}_+^{1\times n}$ is a vector function defined as:

$$\mathrm{Poly}\left(\mathbf{N}, x\right) \equiv \left[1, x, \cdots, x^{n_i-1}\right]_{i=1}^{I}.$$

It is then enough to show that for any $i = 1, 2, \ldots I$,

$$\begin{aligned} &\underbrace{[1, 0, \cdots, 0]}_{1\times n_i} \exp\left(\mathbf{A}_i x\right) = \left[1, x, \cdots, x^{n_i-1}\right] ExpM_i \\ \Longleftrightarrow\ &[1, 0, \cdots, 0] \exp\left(\mathbf{A}_i x\right)\left(ExpM_i\right)^{-1} = \left[1, x, \cdots, x^{n_i-1}\right]. \end{aligned} \tag{A8}$$

Since $ExpM_i$ is a diagonal matrix with identical elements on the diagonal, $\exp\left(\mathbf{A}_i x\right)\left(ExpM_i\right)^{-1}$

becomes:

$$\exp\left(\begin{bmatrix} 0 & 1 & & & \\ & 0 & 2 & & \\ & & 0 & \ddots & \\ & & & \ddots & n_i-1 \\ & & & & 0 \end{bmatrix} x\right) = UPas^x = UPas \circ UToe\left(x\right),$$

where $UPas$ is an n_i-dimensional upper-triangular Pascal matrix (which has a first row of ones), $UToe\left(x\right)$ is an n_i-dimensional upper-triangular Toeplitz matrix of the power series of x:

$$UToe\left(x\right) = \begin{bmatrix} 1 & x & x^2 & \cdots & x^{n_i-1} \\ & 1 & x & \cdots & x^{n_i-2} \\ & & \ddots & & \vdots \\ & & & 1 & x \\ & & & & 1 \end{bmatrix},$$

and $\circ$ denotes the Hadamard product. In fact, $UPas^x$ is also referred to as the transpose of the generalized Pascal matrix of x. For example, see Yang and Micek (2007) and Stefan (2011).

Therefore, the left hand side of (A8) is the first row of the Hadamard product of $UPas$ and $UToe\left(x\right)$, which equals the right hand side of (A8).

D Estimation and forecast of VAR(4) and VARMA(1,1) models

We estimate three-factor VAR and VARMA models, where the three factors are the first three principal components of the zero yields. The estimation for Joslin et al. (2013)'s VAR(4) model is exactly the same as that in JSZ with the number of lags being four instead of one. The estimation for Feunou and Fontaine (2016)'s VARMA(1,1) model is again very similar to that in JSZ, except that the VARMA parameters directly enter the likelihood function and are part of the ML estimation, because they cannot be estimated using OLS regressions. The likelihood function for an unrestricted VARMA(1,1) model can be found in Lütkepohl (2007, p. 464).

Suppose that the VAR(4) model is described by:

$$\mathrm{PC}_t = v + A_1 \mathrm{PC}_{t-\frac{1}{12}} + \cdots + A_4 \mathrm{PC}_{t-\frac{4}{12}} + \sqrt{\Sigma}\epsilon_t,$$

and the VARMA(1,1) model by:

$$(\mathrm{PC}_t - v) = A_1 \left(\mathrm{PC}_{t-\frac{1}{12}} - v\right) + \sqrt{\Sigma}\epsilon_t + M_1 \sqrt{\Sigma}\epsilon_{t-\frac{1}{12}}.$$

Given the parameters estimates, for VAR(4), the h-month ahead forecast is given by:

$$\widehat{\mathrm{PC}}_t \left(\frac{h}{12}\right) = \widehat{v} + \widehat{A}_1 \widehat{\mathrm{PC}}_t \left(\frac{h-1}{12}\right) + \cdots + \widehat{A}_4 \widehat{\mathrm{PC}}_t \left(\frac{h-4}{12}\right).$$

For VARMA(1,1), the h-month ahead forecast is given by:

$$\widehat{\mathrm{PC}}_t \left(\frac{h}{12}\right) = \widehat{v} + \sum_{i=1}^{12(t-1)+h} \widehat{\Pi}_i \left[\widehat{\mathrm{PC}}_t \left(\frac{h-i}{12}\right) - \widehat{v}\right],$$

where $\widehat{\mathrm{PC}}_t\,(j)$ $=\mathrm{PC}_{t+j}$ for $j < 0$, and $\widehat{\Pi}_i = (-1)^{i-1} \left(\widehat{M}_1^i + \widehat{M}_1^{i-1} \widehat{A}_1\right)$.

Denote the model-implied j-year zero yield by:

$$y_t^{(j)} = C_m^{\mathrm{PC}}\left(j\right) + B_m^{\mathrm{PC}}\left(j\right)\mathrm{PC}_t.$$

Then, the model-implied expected excess return is given by:

$$\mathbb{E}_t\left(\overline{rx}_{t+1}\right) = \left(\frac{1}{4}\sum_{j=2}^{5} j y_t^{(j)} - y_t^{(1)}\right) - \frac{1}{4}\sum_{j=1}^{4}\left(jC_m^{\mathrm{PC}}\left(j\right) + jB_m^{\mathrm{PC}}\left(j\right)\widehat{\mathrm{PC}}_t\left(1\right)\right).$$

E Unspanned risk specifications

In this appendix, we show that our framework can be easily extended to cover unspanned risk specifications in the sense of Joslin et al. (2014).

The key to having the unspanned risk property is to allow the stochastic discount factor, hence the market price of risk, to depend on a broader set of state variables, some of which are unspanned by the bond market specific states.[28] Denote these broader state variables by $Z_t^{\mathrm{E}} \in \boldsymbol{R}^N$, which are driven by an M-dimensional Wiener process W_t^{E} under the $\mathbb{P}$ measure. Recall that the bond market specific states $Z_t \subseteq Z_t^{\mathrm{E}}$ are driven by $W_t^{\mathbb{P}} \subseteq W_t^{\mathrm{E}}$, each with dimensionality n and m, respectively, where $n \geq m$, $N \geq n$, $M \geq m$, and $N \geq M$. Since the market price of risk is an affine function of Z_t^{E}, $W_t^{\mathbb{P}}$ is linked to W_t via the following relation:

$$dW_t^{\mathbb{P}} = \left(\underbrace{\lambda_1}_{m\times 1} + \underbrace{\lambda_2}_{m\times N} Z_t^{\mathrm{E}}\right) dt + dW_t. \tag{A9}$$

In light of the recent literature, non-Markov states (Joslin et al., 2013; Feunou and Fontaine, 2016) and additional risk factors, such as macro variables (Joslin et al., 2014), might be unspanned by the bond market specific states. We present two simple examples below to demonstrate how

[28] Joslin et al. (2014) call these broader state variables the "states of the economy."

our framework can accommodate these unspanning features.

A Unspanned non-Markov states

For simplicity, we set $M = m = 1$, $N = 2$, and $n = 1$. Therefore, Z_t under the $\mathbb{Q}$ measure follows:

$$dZ_t = -k_1 Z_t dt + \Omega_1 dW_t. \tag{A10}$$

Since there is no unspanned risk factor in this case, i.e., $W_t = W_t^{\mathrm{E}}$, we specify the dynamics of Z_t^{E} under $\mathbb{Q}$ as:

$$dZ_t^{\mathrm{E}} = \begin{bmatrix} -k_1 & 1 \\ 0 & -k_1 \end{bmatrix} Z_t^{\mathrm{E}} dt + \begin{bmatrix} \Omega_2 \\ \Omega_1 \end{bmatrix} dW_t.$$

Thus the second element in Z_t^{E} is Z_t, i.e., $Z_t^{\mathrm{E}}(2) = Z_t$. As for $Z_t^{\mathrm{E}}(1)$, which captures the lagged information about Z_t, it is unspanned by the current yields because it is excluded from the pricing state variable Z_t.

Given (A9), the $\mathbb{P}$-dynamics of Z_t^{E} is given by:

$$dZ_t^{\mathrm{E}} = -\begin{bmatrix} \Omega_2 \\ \Omega_1 \end{bmatrix} \lambda_1 dt + \left(\begin{bmatrix} -k_1 & 1 \\ 0 & -k_1 \end{bmatrix} - \begin{bmatrix} \Omega_2 \\ \Omega_1 \end{bmatrix} \lambda_2 \right) Z_t^{\mathrm{E}} dt + \begin{bmatrix} \Omega_2 \\ \Omega_1 \end{bmatrix} dW_t^{\mathbb{P}}.$$

B Unspanned additional risk factors

In this example, we assume $M = 2$, $m = 1$, $N = 2$, and $n = 1$. Although Z_t still follows (A10), W_t^{E} includes an additional random source other than $W_t^{\mathbb{P}}$:

$$W_t^{\mathrm{E}} = \begin{bmatrix} W_t^{\mathbb{P}} \\ \\ W_t^{U} \end{bmatrix}.$$

The $\mathbb{P}$-dynamics of Z_t^{E} is specified as:

$$dZ_t^{\mathrm{E}} = \mathbf{A}^{\mathrm{E},\mathbb{P}} \left(Z_t^{\mathrm{E}} - \mu^{\mathrm{E}} \right) dt + \begin{bmatrix} \Omega_1 & 0 \\ \\ \Omega_2 & \Omega_3 \end{bmatrix} dW_t^{\mathrm{E}}.$$

Given the assumption in Joslin et al. (2011) and others about the perfect observability of portfolios of yields and macro variables, the parameters $\mathbf{A}^{\mathrm{E},\mathbb{P}}$, μ^{E}, Ω_1, Ω_2, and Ω_3 can be estimated using standard maximum likelihood. Once they are estimated, the market prices of risk λ_1 and λ_2 are given by the following equations:

$$\begin{aligned} \lambda_1 &= \frac{\left(\mathbf{A}^{\mathrm{E},\mathbb{P}} \mu^{\mathrm{E}}\right)(1)}{\Omega_1}, \\ \lambda_2 &= \frac{\begin{bmatrix} -k_1 & 0 \end{bmatrix} - \mathbf{A}^{\mathrm{E},\mathbb{P}}(1,:)}{\Omega_1}, \end{aligned}$$

where $\left(\mathbf{A}^{\mathrm{E},\mathbb{P}} \mu^{\mathrm{E}}\right)(1)$ and $\mathbf{A}^{\mathrm{E},\mathbb{P}}(1,:)$ denote the first rows of $\mathbf{A}^{\mathrm{E},\mathbb{P}} \mu^{\mathrm{E}}$ and $\mathbf{A}^{\mathrm{E},\mathbb{P}}$, respectively.

References

Ahn, D.-H. and B. Gao (1999). A parametric nonlinear model of term structure dynamics. *Review of financial studies 12*(4), 721–762.

Ang, A. and M. Piazzesi (2003). A no-arbitrage vector autoregression of term structure dynamics with macroeconomic and latent variables. *Journal of Monetary economics 50*(4), 745–787.

Athanasopoulos, G. and F. Vahid (2008). Varma versus var for macroeconomic forecasting. *Journal of Business & Economic Statistics 26*(2).

Barberis, N., R. Greenwood, L. Jin, and A. Shleifer (2015). X-capm: An extrapolative capital asset pricing model. *Journal of Financial Economics 115*(1), 1 – 24.

Bauer, M. D. and G. D. Rudebusch (2015). Resolving the spanning puzzle in macro-finance term structure models. Technical report, Federal Reserve Bank of San Francisco Working Paper Series.

Bauer, M. D., G. D. Rudebusch, and J. C. Wu (2012). Correcting estimation bias in dynamic term structure models. *Journal of Business & Economic Statistics 30*(3), 454–467.

Bergstrom, A. R. (1983). Gaussian estimation of structural parameters in higher order continuous time dynamic models. *Econometrica: Journal of the Econometric Society*, 117–152.

Björk, T. and A. Gombani (1999). Minimal realizations of interest rate models. *Finance and Stochastics 3*(4), 413–432.

Björk, T. and L. Svensson (2001). On the Existence of Finite-Dimensional Realizations for Nonlinear Forward Rate Models. *Mathematical Finance 11*(2), 205–243.

Brace, A. and M. Musiela (1994). A MULTIFACTOR GAUSS MARKOV IMPLEMENTATION OF HEATH, JARROW, AND MORTON. *Mathematical Finance 4*(3), 259–283.

Chen, B. and Y. Hong (2011). Testing for the markov property in time series. *Econometric Theory 28*(1), 130.

Chiarella, C. and O. Kwon (2003). Finite Dimensional Affine Realisations of HJM Models in Terms of Forward Rates and Yields. *Review of Derivatives Research 6*(2), 129–155.

Cieslak, A. and P. Povala (2014). Expecting the fed. *Available at SSRN 2239725*.

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

Cochrane, J. H. and M. Piazzesi (2006). Appendix to "bond risk premia". *Chicago Booth and Standford GSB*.

Collin-Dufresne, P., R. Goldstein, and C. Jones (2008). Identification of maximal affine term structure models. *The Journal of Finance 63*(2), 743–795.

Cox, J., J. Ingersoll Jr, and S. Ross (1985). A theory of the term structure of interest rates. *Econometrica: Journal of the Econometric Society 53*(2), 385–407.

Cox, J. C., J. E. Ingersoll, and S. A. Ross (1981). A re-examination of traditional hypotheses about the term structure of interest rates. *The Journal of Finance 36*(4), 769–799.

Dai, Q. and K. Singleton (2000). Specification analysis of affine term structure models. *Journal of Finance 55*(5), 1943–1978.

Dai, Q. and K. Singleton (2003). Term structure dynamics in theory and reality. *Review of Financial Studies 16*(3), 631–678.

Dai, Q., K. J. Singleton, and W. Yang (2007). Regime shifts in a dynamic term structure model of us treasury bond yields. *Review of Financial Studies 20*(5), 1669–1706.

De Jong, F. and P. Santa-Clara (1999). The Dynamics of the Forward Interest Rate Curve: A Formulation with State Variables. *Journal of Financial and Quantitative Analysis 34*(1), 131–157.

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

Duffee, G. R. (2011). Information in (and not in) the term structure. *Review of Financial Studies 24*(9), 2895–2934.

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

Dumas, B., A. Kurshev, and R. Uppal (2009). Equilibrium portfolio strategies in the presence of sentiment risk and excess volatility. *The Journal of Finance 64*(2), 579–629.

Evans, C. L. and D. A. Marshall (1998). Monetary policy and the term structure of nominal interest rates: evidence and theory. In *Carnegie-Rochester Conference Series on Public Policy*, Volume 49, pp. 53–111. Elsevier.

Evans, C. L. and D. A. Marshall (2007). Economic determinants of the nominal treasury yield curve. *Journal of Monetary Economics 54*(7), 1986–2003.

Feunou, B. and J.-S. Fontaine (2014). Non-markov gaussian term structure models: The case of inflation. *Review of Finance 18*(5), 1953–2001.

Feunou, B. and J.-S. Fontaine (2016). Gaussian term structure models and bond risk premia. *Forthcoming in Management Science*.

Gourieroux, C., A. Monfort, F. Pegoraro, and J.-P. Renne (2014). Regime switching and bond pricing. *Journal of Financial Econometrics 12*(2), 237–277.

Heath, D., R. Jarrow, and A. Morton (1992). Bond Pricing and the Term Structure of Interest Rates: A New Methodology for Contingent Claims Valuation. *Econometrica 60*(1), 77–105.

Higham, N. J. (2008). *Functions of matrices: theory and computation*. Siam.

Jardet, C., A. Monfort, and F. Pegoraro (2013). No-arbitrage Near-Cointegrated VAR(p) term structure models, term premia and GDP growth. *Journal of Banking & Finance 37*(2), 389 – 402.

Jin, Y. and P. Glasserman (2001). Equilibrium positive interest rates: a unified view. *Review of Financial Studies 14*(1), 187–214.

Joslin, S., A. Le, and K. J. Singleton (2013). Gaussian macro-finance term structure models with lags. *Journal of Financial Econometrics 11*(4), 581–609.

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

Joslin, S., K. Singleton, and H. Zhu (2011). A new perspective on gaussian dynamic term structure models. *Review of Financial Studies*.

Li, H., T. Li, and C. Yu (2013). No-arbitrage taylor rules with switching regimes. *Management Science 59*(10), 2278–2294.

Longstaff, F. A. and E. S. Schwartz (1992). Interest rate volatility and the term structure: A two-factor general equilibrium model. *The Journal of Finance 47*(4), 1259–1282.

Lütkepohl, H. (2007). *New introduction to multiple time series analysis*. Springer.

Merton, R. C. (1973). An intertemporal capital asset pricing model. *Econometrica: Journal of the Econometric Society*, 867–887.

Monfort, A. and F. Pegoraro (2007). Switching varma term structure models. *Journal of Financial Econometrics 5*(1), 105–153.

Stefan, S. (2011). A generalization of the pascal matrix and its properties. *Facta universitatis-series: Mathematics and Informatics* (26), 17–27.

Trolle, A. and E. Schwartz (2009). A general stochastic volatility model for the pricing of interest rate derivatives. *Review of Financial Studies 22*(5), 2007.

Vasicek, O. (1977). An equilibrium characterization of the term structure. *Journal of Financial Economics 5*(2), 177–188.

Wright, J. H. (2011). Term premia and inflation uncertainty: Empirical evidence from an international panel dataset. *The American Economic Review 101*(4), 1514–1534.

Wymer, C. R. (1993). Continuous-time models in macroeconomics: specification and estimation. In *Continuous-Time Econometrics*, pp. 35–79. Springer.

Yang, Y. and C. Micek (2007). Generalized pascal functional matrix and its applications. *Linear algebra and its applications 423*(2-3), 230–245.