Supply and Demand and the Term Structure of Interest Rates *

Robin Greenwood   Samuel Hanson   Dimitri Vayanos

October 2023

Abstract

We survey the growing literature emphasizing the role that supply-and-demand forces play in shaping the term structure of interest rates. Our starting point is the Vayanos and Vila (2009, 2021) model of the term structure of default-free bond yields, which we present in both discrete and continuous time. The key friction in the model is that the bond market is partially segmented from other financial markets: the prices of short-rate and bond supply risk are set by specialized bond arbitrageurs who must absorb shocks to the supply and demand for bonds from other "preferred-habitat" agents. We discuss extensions of this model in the context of default-free bonds and other asset classes.

JEL Codes: E43, E52, F31, G12

*Greenwood and Hanson are from the Harvard Business School and the NBER; Vayanos is from the London School of Economics, the NBER, and CEPR. Disclosures are available on the authors' websites.

1 Introduction

The past 15 years have witnessed a resurgence of interest in the role that supply-and-demand forces play in shaping the term structure of interest rates. In part, this interest has been driven by the widespread adoption of quantitative easing (QE) policies by major central banks which, beginning in 2008, have sought to depress long-term interest rates by purchasing vast quantities of long-term bonds. An enormous literature has demonstrated that large shocks to the supply of long-term bonds—such as those stemming from QE—and shocks to demand from investors such as pension funds exert significant effects on long-term interest rates. For example, Gagnon et al. (2011) report that the combined announcement effect of U.S. QE policies between 2008 and 2010 was to reduce 10-year U.S. Treasury yields by 91 basis points. Similar effects have been documented in numerous other countries, including the U.K, Japan, and the Eurozone.[1]

The role of supply-and-demand forces in the term structure is puzzling, both from the perspective of textbook macroeconomic theory and of traditional asset-pricing theory. In textbook macroeconomic models with intergenerational risk sharing and non-distortionary taxes, Ricardian equivalence holds. As a result, households' consumption profiles and the term structure of interest rates do not depend on whether the government finances its expenditures using debt or taxes, or whether it borrows using short-term or long-term bonds. For instance, consider a government that seeks to reduce long-term interest rates by buying back long-term bonds and replacing them with short-term bonds—i.e., by pursuing a form of QE. Because households sell long-term bonds to the government, they will realize smaller capital losses if interest rates rise in the future. However, because the government's outstanding debt has become shorter term, households must pay higher taxes if interest rates rise. These two effects exactly offset each other, leaving households' consumption profiles and interest rates unchanged. As a result, the maturity structure of the government's debt is irrelevant. Eggertsson and Woodford (2003) prove irrelevance propositions along these lines.

The role of supply-and-demand forces in the term structure is also puzzling from the perspective of traditional asset-pricing theory, which assumes frictionless and fully integrated financial markets. In models with fully integrated financial markets, changes in the net supply of debt, whether short- or long-term, can only impact bond risk premia and yields through their impact on the aggregate risks that highly diversified investors must bear in equilibrium and only to the extent to which these risks are correlated with investors' marginal utility. But if global financial markets are fully integrated, even seemingly large supply shocks such as those stemming from QE policies are a tiny drop in the ocean of global aggregate risk. This explains why Federal Reserve chair Benjamin Bernanke once quipped "The problem with Quantitative Easing (QE) is that it works in practice but not in theory" (Bernanke 2014).

[1] Table 1 in Wiliams (2014) collects empirical estimates from studies of QE in the U.S. and other countries.

A more promising and natural approach to understanding how changes in the supply and demand for bonds affect the term structure has been to assume that the investors who must accommodate these changes are not the highly diversified investors envisioned by traditional asset-pricing theory but instead are bond market specialists. For these specialized investors to accommodate changes in the supply and demand for bonds, bond risk premia must adjust to compensate them for the additional risk they are bearing. This approach to bond markets was first proposed by Tobin (1958, 1969). Modigliani and Sutch (1966) later introduced the notion that investors differ in their "preferred habitats" along the term structure, with some investors preferring long maturities and some preferring shorter ones. This approach has been modernized and substantially generalized by Vayanos and Vila (2009, 2021, hereafter VV), generating a host of additional predictions.

In the VV model, the short-term interest rate, and hence the expectations hypothesis (EH) component of longer-term bond yields, follows an exogenous stochastic process. The key focus is on how bond risk premia—the expected excess returns on longer-term bonds over the short rate—are shaped by supply and demand in equilibrium. The key friction in the VV model is that the bond market is partially segmented from broader financial markets as well as from the economy at large. Formally, the marginal investors in bonds—whom we call "bond arbitrageurs"—are specialized traders who choose portfolios consisting of short- and long-term bonds. Arbitrageurs have mean-variance preferences over next-period wealth and accommodate changes in the supply and demand for bonds from other agents. Changes in bond supply and demand can arise from government issuance, from central bank purchases, or from other investors such as pension funds who have a demand for long-term bonds. We refer to all bond market participants other than arbitrageurs as "preferred-habitat agents." We refer to the net supply coming from these preferred-habitat agents as "supply," although this should be understood as gross supply from issuers minus demand from preferred-habitat investors. Since specialized bond arbitrageurs are risk-averse, they will only absorb supply shocks from preferred-habitat agents if bond expected returns adjust in response.

Over the 15 years, the VV approach has become the standard model for understanding the impact of large-scale bond purchases by central banks, as well as a host of other supply-and-demand driven phenomena in bond markets. The VV model has also been extended to make sense of similar supply-and-demand effects in other asset classes, including foreign exchange, interest-rate derivatives, and credit instruments.

In this review, we present the VV model of the term structure and its applications and extensions. While the original VV model was developed in continuous time, we begin by presenting it in discrete time, as was first done by Hanson (2014). We then present the continuous-time version of the VV model in parallel, and briefly discuss a simplified version in which there are only short-term bonds and a single class of perpetual long-term bonds. Our goal is to present a set of workhorse models that researchers can adapt for use in other settings and to sketch out a number of ways that researchers have already adapted the VV model. We both draw out the key economic intuitions that emerge from these models and present the "cookbook" for solving them.

We begin our analysis by studying a setting where the only risky assets are a set of zero-coupon bonds that are present in constant and inelastic supply. Bond arbitrageurs are risk averse and can flexibly allocate their wealth amongst bonds of different maturities. Critically, "bad times" for specialized bond arbitrageurs need not coincide with bad times for well-diversified investors or for the representative household. Specifically, under the natural assumption that specialized bond arbitrageurs typically have a long position in long-term bonds, periods of rising interest rates will be "bad times" for bond arbitrageurs, and long-term bonds must offer a positive risk premium even though rising interest rates often imply good news for well-diversified investors and the overall macroeconomy. Thus, in contrast to traditional integrated-market theories, the VV approach makes it easy to understand why bond term premia are usually positive and why the yield curve is typically upward sloping. Specifically, if bond supply is positive, the yield curve is upward sloping on average, and its average slope is proportional to the amount of dollar-duration risk that arbitrageurs must bear, arbitrageur risk aversion, and the variance of short-rate shocks.

With a single risk factor, the effects that any supply shock has on the term structure work through the shock's contribution to dollar-duration risk. To illustrate the striking implications of this result, we use Figure 1, first introduced by Cochrane (2008). Suppose that there is an unanticipated and permanent supply shock consisting of an increase in the supply of 10-year bonds and an equal dollar reduction in the supply of 3-year bonds. What happens to the yield curve? A naïve but incorrect intuition is that the effects are localized, with 10-year yields rising and 3-year yields declining. But this intuition misses a central insight from the model: with a single risk factor—i.e., the short-term interest rate—bond returns are perfectly correlated across maturities. As a result, the net impact of this shock depends solely on how it alters arbitrageurs' overall exposure to the short rate. Specifically, since the price of the 10-year bond is more sensitive to movements in the short rate than that of the 3-year bond, the supply shock raises the dollar-duration risk that arbitrageurs must bear, thereby pushing up the equilibrium price of short-rate risk. As a result, the supply shock leads the entire yield curve to steepen. In summary, the single-factor VV model implies that local supply-and-demand shocks have global effects on the yield curve.

We next study the case when the supply of zero-coupon bonds responds elastically to bond prices. We consider the case where the supply curve is upward-sloping as well as the case where it is downward-sloping. Upward-sloping bond supply could arise if the government and firms tend to issue more debt when interest rates are low, or if bond demand from preferred-habitat investors is downward-sloping. We show that upward-sloping bond supply implies that bond risk premia will be low when short rates are high, leading long-term yields to underreact to movements in short rates relative to the EH benchmark. Thus, when bond supply is upward-sloping, the VV model can match the stylized facts from Fama (1984)

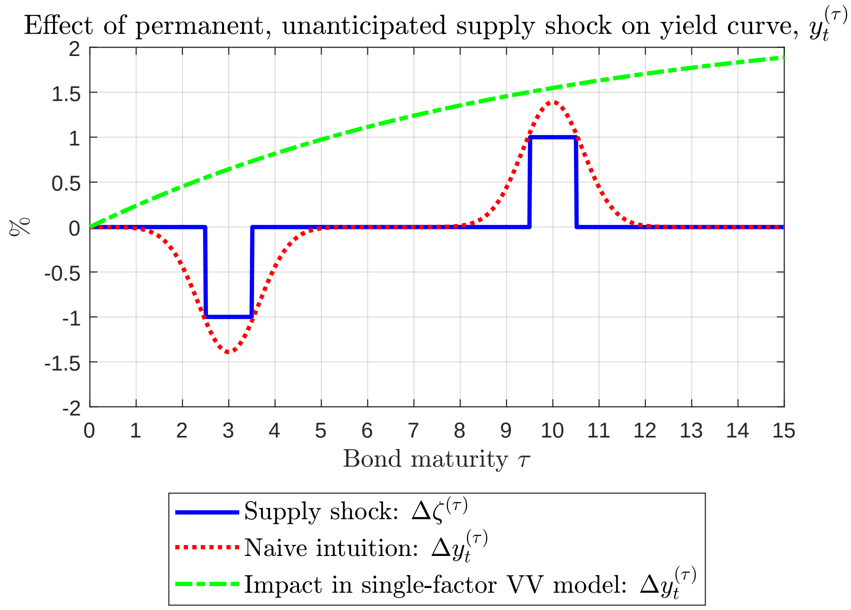

Figure 1: Effects of bond supply with one risk factor.

Campbell and Shiller (1991) that a steeper yield curve is associated with higher bond risk premia.

Perhaps surprisingly, at higher frequencies bond supply may be downward-sloping because of mortgage refinancing or extrapolative investors. We show that a downward-sloping bond supply generates a term premium that is high when short rates are high, leading bond yields to overreact to movements in short rates relative to the EH. As a result, VV models with downward-sloping supply have been used to explain the finding that long-term rates appear to be excessively sensitive to movements in short rates at higher frequencies.

We next modify the model to consider the case where there are random shocks to bond supply from preferred-habitat agents as in Greenwood and Vayanos (2014). Supply risk creates an additional source of priced risk that is reflected in the term structure. A key finding in this case is that the expected persistence of supply shocks determines their impact on the term structure of interest rates. If supply shocks are highly persistent, then they will have a monotonic effect on the yield curve, with the greatest effect being on long-term yields. However, if supply shocks are more transient, then they will have a hump-shaped effect on the yield curve, with the greatest effect being on intermediate-term yields.

The model we just described is developed in Section 2. In Section 3 we repeat a portion of the analysis in continuous time. Wherever possible, we use identical notation so that researchers can seamlessly switch between the two presentations. While our discrete-time model relies on an approximation that becomes exact only in the continuous-time limit, the derivations and formulas in the discrete-time model have perfect analogs in continuous time.

But, as we show, the model is straightforward to express in either setting, with the discrete-time setting being simpler for some applications and the continuous-time setting for others. The advantage of discrete time is that it may be easier to incorporate into other models.

The continuous-time version of the model turns out to be well suited, however, for describing the case where supply is both random and price-elastic. When supply is both random and elastic, even permanent supply shocks can have hump-shaped effects. The intuition for the hump shape is that because arbitrageurs sell bonds at a given maturity to accommodate a decrease in supply, they become more willing to buy bonds at other maturities to keep a similar risk profile. With elastic supply, the prices of these other bonds move only to the extent that arbitrageurs actually buy them. Arbitrageurs are not willing to buy them aggressively because, with random supply, bond returns are imperfectly correlated across maturities.

Finally, in Section 4 we develop a simplified version of the VV model in which there are only short-term bonds and a single class of perpetual long-term bonds. While this model requires an additional approximation, it is very easy to solve. As a result, it may be useful to macro-economists or macro-finance researchers who want to model the idea that supply-and-demand effects play a role shaping long-term interest rates, but who are not interested in modeling the full term structure of yields.

Since the first version of VV was circulated in 2007, the theoretical and empirical literature on supply-and-demand effects in the term structure has blossomed. In part, this growth has been driven by the widespread interest in QE policies. In Section 5 we provide a brief overview of empirical applications of the VV framework, with the objective of connecting the findings to key economic ideas highlighted by the VV model. Most empirical applications of the VV framework have either sought to isolate shocks to the supply or demand for long-term bonds (such as QE or demand by pension funds), or alternately sought to construct empirical proxies for arbitrageurs' holdings of long-term bonds and risk aversion.

In Sections 6 and 7 we review several theoretical extensions of VV that have been developed in literature. Section 6 focuses on extensions that focus on the term structure of default-free bond yields. The main extensions that we review are multiple supply factors (Vayanos and Vila (2009, 2021)), forward guidance about short rates and bond supply (Greenwood et al. (2016)), the zero-lower bound on short rates (King (2019)), slow-moving capital (Greenwood et al. (2018)), arbitrageur wealth effects and balance-sheet frictions (Kekre et al. (2023),He et al. (2022),Hanson et al. (2023)), convenience yields that arise because government bonds have some of the valuable attributes of money (Krishnamurthy and Vissing-Jorgensen (2012)), and real versus nominal yields (Campbell et al. (2009)).

Section 7 discusses applications to other domains, including foreign exchange (Greenwood et al. (2023), Gourinchas et al. (2022)), defaultable corporate bonds (Greenwood et al. (2018),Costain et al. (2022)), and investor segmentation across markets. Section 8 concludes, with some remarks about directions for future research.

2 The VV Model in Discrete Time

2.1 Model

Time t is discrete and infinite. At each time t, there are zero-coupon, default-free bonds that have face value one and mature in $\tau = 1, 2, ..., T$ periods. We refer to the bond maturing in τ periods as the τ-period bond. The price of the τ-period bond at time t is denoted by

$$P_t^{(\tau)} = \exp(-\tau y_t^{(\tau)}), \tag{1}$$

where $y_t^{(\tau)}$ denotes the bond's continuously compounded yield to maturity. We denote the log bond price by $p_t^{(\tau)} \equiv \log(P_t^{(\tau)}) = -\tau y_t^{(\tau)}$.

We refer to the yield of the 1-period bond as the short rate and denote it by $r_t = y_t^{(1)}$. We take r_t as exogenous and assume that it follows the AR(1) process

$$r_{t+1} = \overline{r} + \rho_r(r_t - \overline{r}) + \sigma_r \varepsilon_{r,t+1}, \tag{2}$$

where $\rho_r \in (0,1)$, $\overline{r}$, and $\sigma_r > 0$ are constants, and $\varepsilon_{r,t+1}$ is a stochastic shock with $\mathbb{E}_t[\varepsilon_{r,t+1}] = 0$ and $\mathbb{V}\text{ar}_t[\varepsilon_{r,t+1}] = 1$. In the background, we think of monetary policy as determining the short rate r_t outside of the model.

We denote the return on the τ-period bond between times t and $t+1$ by $R_{t+1}^{(\tau)} \equiv P_{t+1}^{(\tau-1)}/P_t^{(\tau)} - 1$, and the log return by $r_{t+1}^{(\tau)} \equiv \log(1 + R_{t+1}^{(\tau)}) = p_{t+1}^{(\tau-1)} - p_t^{(\tau)}$. We denote these returns by $R_{t+1}^{(1)} = 1/P_t^{(1)} - 1 \equiv R_t$ and $r_{t+1}^{(1)} = -p_t^{(1)} = r_t$, respectively, for the 1-period bond. The return on the τ-period bond between times t and $t+1$ can be approximated by

$$R_{t+1}^{(\tau)} \approx r_{t+1}^{(\tau)} + \frac{1}{2}\mathbb{V}\text{ar}_t[r_{t+1}^{(\tau)}]. \tag{3}$$

The approximation becomes exact in the continuous-time limit of Section 3 as the time between periods goes to zero.[2]

There are two types of agents, "bond arbitrageurs" and "preferred-habitat agents". Arbitrageurs are competitive and form a continuum with unit mass. They choose their bond portfolios to trade off the mean and variance of their wealth 1-period ahead. Denoting by $X_t^{(\tau)}$ the market value of arbitrageurs' holdings of the τ-period bond at time t, arbitrageur

[2]This approximation follows from the second-order Taylor expansion

$$1 + R_{t+1}^{(\tau)} = \exp(r_{t+1}^{(\tau)}) \approx 1 + r_{t+1}^{(\tau)} + \frac{1}{2}(r_{t+1}^{(\tau)})^2 = 1 + r_{t+1}^{(\tau)} + \frac{1}{2}\left((r_{t+1}^{(\tau)} - \mathbb{E}_t[r_{t+1}^{(\tau)}]) + \mathbb{E}_t[r_{t+1}^{(\tau)}]\right)^2,$$

and because, in the vicinity of the continuous-time limit, the term $(r_{t+1}^{(\tau)} - \mathbb{E}_t[r_{t+1}^{(\tau)}])^2$ is close to $\mathbb{V}\text{ar}_t[r_{t+1}^{(\tau)}]$, and the terms $2(r_{t+1}^{(\tau)} - \mathbb{E}_t[r_{t+1}^{(\tau)}])\mathbb{E}_t[r_{t+1}^{(\tau)}]$ and $(\mathbb{E}_t[r_{t+1}^{(\tau)}])^2$ are small relative to $r_{t+1}^{(\tau)}$ and $\mathbb{V}\text{ar}_t[r_{t+1}^{(\tau)}]$. Using log returns plus a Jensen's inequality adjustment to approximate simple net returns as in (3) is a linearity-generating modeling device that does not qualitatively impact any of our conclusions. Indeed, the formulae we derive below are the exact discrete-time analogs of the formulae that arise in the continuous-time limit.

wealth evolves according to

$$W_{t+1} = (W_t - \sum\nolimits_{\tau=2}^{T} X_t^{(\tau)})(1+R_t) + \sum\nolimits_{\tau=2}^{T} X_t^{(\tau)}(1+R_{t+1}^{(\tau)}). \tag{4}$$

The first term in (4) is arbitrageurs' return from investing in the short rate, and the second term is the return from investing their remaining wealth in τ-period bonds for $\tau \geq 2$. Using (3), we can approximate the evolution of arbitrageurs' wealth by

$$W_{t+1} \approx \hat{W}_{t+1} \equiv W_t(1+r_t) + \sum\nolimits_{\tau=2}^{T} X_t^{(\tau)}(r_{t+1}^{(\tau)} + \frac{1}{2}\mathbb{V}\mathrm{ar}_t[r_{t+1}^{(\tau)}] - r_t). \tag{5}$$

We assume that arbitrageurs choose their bond holdings $\{X_t^{(\tau)}\}_{\tau=2}^{T}$ to maximize

$$\mathbb{E}_t[\hat{W}_{t+1}] - \frac{a}{2}\mathbb{V}\mathrm{ar}_t[\hat{W}_{t+1}], \tag{6}$$

where $a \geq 0$ is arbitrageurs' coefficient of risk aversion.

We use the term "preferred-habitat agents" to refer to all bond market participants other than arbitrageurs, including the government and other players who issue bonds as well as other investors who hold bonds. To clear the market, arbitrageurs must hold the net supply of bonds coming from preferred-habitat agents. For simplicity, we refer to this net supply as "supply." However, it is worth bearing in mind that shifts in supply can stem either from shifts in the amount of outstanding bonds—e.g., due to issuance by the government—or from shifts in the holdings of other investors. We model supply in reduced from and consider different specifications for it in Sections 2.2 to 2.4.

We end this section by deriving a general relationship between bond yields, short rates, and bond excess returns that holds in all the equilibria that we derive in Sections 2.2 to 2.4. The return on the τ-period bond between times t and $t+1$ can be written as the sum of the expected component $\mathbb{E}_t[R_t^{(\tau)}]$ and the unexpected component $R_t^{(\tau)} - \mathbb{E}_t[R_t^{(\tau)}]$. Using (3), we can write the expected and unexpected components as

$$\mu_t^{(\tau)} \equiv \mathbb{E}_t[r_{t+1}^{(\tau)}] + \frac{1}{2}\mathbb{V}\mathrm{ar}_t[r_{t+1}^{(\tau)}] = \mathbb{E}_t[p_{t+1}^{(\tau-1)} - p_t^{(\tau)}] + \frac{1}{2}\mathbb{V}\mathrm{ar}_t[p_{t+1}^{(\tau-1)}], \tag{7}$$

$$r_{t+1}^{(\tau)} - \mu_t^{(\tau)} = r_{t+1}^{(\tau)} - \mathbb{E}_t[r_{t+1}^{(\tau)}] = p_{t+1}^{(\tau-1)} - \mathbb{E}_t[p_{t+1}^{(\tau-1)}], \tag{8}$$

respectively. Using (7), the fact that $p_t^{(\tau)} = -\tau y_t^{(\tau)}$, and assuming that $\mathbb{V}\mathrm{ar}_t[r_{t+1}^{(\tau)}]$ is constant over time as will be the case in the equilibrium of our model, we can write the expected excess return over the short rate as

$$\mu_t^{(\tau)} - r_t = \tau y_t^{(\tau)} - (\tau-1)\mathbb{E}_t[y_{t+1}^{(\tau-1)}] - r_t + \frac{1}{2}\mathbb{V}\mathrm{ar}^{(\tau)}, \tag{9}$$

where $\mathbb{V}\text{ar}^{(\tau)}$ denotes the constant value of $\mathbb{V}\text{ar}_t[r_{t+1}^{(\tau)}]$. Iterating (9) forward, we find

$$y_t^{(\tau)} = \underbrace{\tau^{-1}\sum\nolimits_{j=0}^{\tau-1}\mathbb{E}_t\left[r_{t+j}\right]}_{\text{Expected short rates}} + \underbrace{\tau^{-1}\sum\nolimits_{j=0}^{\tau-1}\mathbb{E}_t[\mu_{t+j}^{(\tau-j)} - r_{t+j}]}_{\text{Expected excess bond returns}} - \underbrace{\tau^{-1}\sum\nolimits_{j=0}^{\tau-1}\frac{1}{2}\mathbb{V}\text{ar}^{(\tau-j)}}_{\text{Convexity adjustment}}. \tag{10}$$

The yield of the τ-period bond is equal to the sum of three terms. The first term is average expected short rates over the bond's life. This term corresponds to the Expectations Hypothesis (EH) component of the term structure. The second term is the average of the bond's 1-period expected returns in excess of the short rate over the bond's life. The third term is a convexity adjustment. Since expected excess returns in our model arise purely from risk, we refer to them as risk premia from now on.

2.2 Constant Supply

In this section we assume that the supply coming from preferred-habitat agents, expressed in market-value terms, is constant over time.[3] Denoting the supply of the τ-period bond in period t by $S_t^{(\tau)}$, we assume

$$S_t^{(\tau)} = \zeta^{(\tau)} \tag{11}$$

for a function $\zeta^{(\tau)}$ that depends only on maturity τ.

We look for an equilibrium where bond yields are affine functions of the short rate. In our conjectured equilibrium, the log price of the τ-period bond at time t takes the form

$$p_t^{(\tau)} = -\tau y_t^{(\tau)} = -\left[A_r^{(\tau)}(r_t - \overline{r}) + C^{(\tau)}\right], \tag{12}$$

where the functions $A_r^{(\tau)}$ and $C^{(\tau)}$ depend only on maturity τ. The function $A_r^{(\tau)}$ characterizes the sensitivity of bond prices to the short rate factor r_t as a function of bond maturity τ, where we define factor sensitivity or dollar duration with respect to the factor as percentage decline in price per unit increase in the factor.[4]

We solve for the equilibrium when supply is constant in Section 2.2.1. Section 2.2.2 presents the key results and intuitions and is mostly self-contained. Thus, readers who want to skip the mathematical derivations in Section 2.2.1 can skip ahead to Section 2.2.2.

[3]Expressing supply in market-value terms—as opposed to face-value terms—is key to the tractability of the VV model. Although this assumption is not completely innocuous, broadly similar economic conclusions would hold if supply was instead specified in face-value terms.

[4]Since $p_t^{(\tau)} = -\tau y_t^{(\tau)}$, the *duration* of the τ-period bond—i.e., its percentage decline in price per unit increase in its yield—is trivially τ. However, from (12), the *dollar duration* of the τ-period bond with respect to the short rate r_t is $A_r^{(\tau)}$.

2.2.1 Deriving the Equilibrium

Substituting (12) into (7) and (8), and using the assumed AR(1) dynamics for the short rate, we can write the expected and unexpected components of the return of the τ-period bond between times t and $t+1$ as

$$\mu_t^{(\tau)} = A_r^{(\tau)}(r_t - \overline{r}) - A_r^{(\tau-1)}\rho_r(r_t - \overline{r}) + C^{(\tau)} - C^{(\tau-1)} + \frac{\sigma_r^2}{2}[A_r^{(\tau-1)}]^2, \tag{13}$$

$$r_{t+1}^{(\tau)} - \mu_t^{(\tau)} = -A_r^{(\tau-1)}\sigma_r\varepsilon_{r,t+1}, \tag{14}$$

respectively. Substituting (13) and (14) into arbitrageurs' budget constraint (5), we can write the arbitrageurs' objective (6) as

$$\sum\nolimits_{\tau=2}^{T} X_t^{(\tau)}(\mu_t^{(\tau)} - r_t) - \frac{a\sigma_r^2}{2}\left(\sum\nolimits_{\tau=2}^{T} X_t^{(\tau)} A_r^{(\tau-1)}\right)^2. \tag{15}$$

Arbitrageurs maximize (15) over their bond holdings $\{X_t^{(\tau)}\}_{\tau=2}^{T}$. The arbitrageurs' first-order condition for their holdings of the τ-period bond, $X_t^{(\tau)}$, is

$$\mu_t^{(\tau)} - r_t = A_r^{(\tau-1)}\lambda_{r,t}, \tag{16}$$

where

$$\lambda_{r,t} \equiv a\sigma_r^2 \sum\nolimits_{\tau=2}^{T} X_t^{(\tau)} A_r^{(\tau-1)} \tag{17}$$

is the price of short-rate risk.

The first-order condition (16) follows from the absence of arbitrage. Specifically, absence of arbitrage requires that the ratio of a bond's risk premium $\mu_t^{(\tau)} - r_t$ to the bond's sensitivity $A_r^{(\tau-1)}$ to the short-rate risk factor must be the same for all bonds. If the ratio differed across bonds, then it would be possible to construct riskless arbitrage portfolios. However, the absence of arbitrage alone imposes no restrictions on the price of short-rate risk $\lambda_{r,t}$.[5]

We determine the equilibrium price of short-rate risk $\lambda_{r,t}$ from market-clearing—i.e., from the equilibrium interaction between arbitrageurs and preferred-habitat agents. Substituting arbitrageurs' position $X_t^{(\tau)}$ from the market-clearing condition $X_t^{(\tau)} = S_t^{(\tau)} = \zeta^{(\tau)}$ into (17), we find

$$\lambda_{r,t} = a\sigma_r^2 \sum\nolimits_{\tau=2}^{T} \zeta^{(\tau)} A_r^{(\tau-1)} \equiv \lambda_r. \tag{18}$$

The price of short-rate risk is proportional to arbitrageurs' risk aversion a, the variance σ_r^2 of short-rate shocks, and the sensitivity $\sum_{\tau=2}^{T} \zeta^{(\tau)} A_r^{(\tau-1)}$ of arbitrageurs' bond portfolio to the short-rate factor. The portfolio's factor sensitivity or dollar duration with respect to the short-rate factor is obtained by multiplying the arbitrageurs' position $\zeta^{(\tau)}$ in the τ-period

[5] Technically, (16) only follows *approximately* from the absence of arbitrage since (13) and (14) only hold approximately in discrete time. However, this equation follows strictly from the absence of arbitrage in the model's continuous-time limit since the continuous-time analogs to (13) and (14) hold exactly.

bond by the bond's sensitivity $A_r^{(\tau-1)}$ to the factor and then summing across maturities. The price of short-rate risk is constant over time and is denoted by λ_r, because the constant supply assumption implies that arbitrageurs' bond positions are constant.

Substituting $\mu_t^{(\tau)}$ from (13) and λ_r from (18) into (16), we obtain the equation

$$A_r^{(\tau)}(r_t-\overline{r})-A_r^{(\tau-1)}\rho_r(r_t-\overline{r})+C^{(\tau)}-C^{(\tau-1)}+\frac{\sigma_r^2}{2}[A_r^{(\tau-1)}]^2-r_t=A_r^{(\tau-1)}\lambda_r, \tag{19}$$

which we use to solve for the functions $A_r^{(\tau)}$ and $C^{(\tau)}$. Equation (19) is affine in $r_t-\overline{r}$. Identifying linear terms in $r_t-\overline{r}$, we find the difference equation

$$A_r^{(\tau)}-A_r^{(\tau-1)}\rho_r-1=0. \tag{20}$$

Identifying constant terms, we find the difference equation

$$C^{(\tau)}-C^{(\tau-1)}-\overline{r}-A_r^{(\tau-1)}\lambda_r+\frac{\sigma_r^2}{2}[A_r^{(\tau-1)}]^2=0. \tag{21}$$

The terminal conditions for (20) and (21) are $A_r^{(0)}=0$ and $C^{(0)}=0$, respectively, because maturing bonds are worth their face value of one (i.e., $P_t^{(0)}=1$).

The solution for $A_r^{(\tau)}$ is

$$A_r^{(\tau)}=\underbrace{\sum\nolimits_{j=0}^{\tau-1}(\rho_r)^j}_{\frac{\partial}{\partial r_t}\left(\sum_{j=0}^{\tau-1}\mathbb{E}_t[r_{t+j}]\right)}=\frac{1-\rho_r^\tau}{1-\rho_r}, \tag{22}$$

and the solution for $C^{(\tau)}$ is

$$C^{(\tau)}=\underbrace{\tau\overline{r}}_{\sum_{j=0}^{\tau-1}\mathbb{E}[r_{t+j}]}+\underbrace{\left[\sum\nolimits_{j=1}^{\tau}A_r^{(j-1)}\right]\lambda_r}_{\sum_{j=0}^{\tau-1}\mathbb{E}[\mu_{t+j}^{(\tau-j)}-r_{t+j}]}-\underbrace{\frac{\sigma_r^2}{2}\sum\nolimits_{j=1}^{\tau}[A_r^{(j-1)}]^2}_{\sum_{j=0}^{\tau-1}\mathbb{V}\mathrm{ar}^{(\tau-j)}}. \tag{23}$$

Note that $\tau^{-1}C^{(\tau)}$ is the yield of the τ-period bond when the short rate r_t is equal to its long-run mean $\overline{r}$. According to (23), that yield is equal to the sum of three terms which correspond to those in the yield decomposition (10). The first term, $\overline{r}$, is the average of the expected short rates over the next τ periods. The second term is the average of the risk premia earned by the bond (and given by (16)) over the next τ periods. The third term is a convexity adjustment.

2.2.2 Results and Intuitions

In this section we examine how shocks to the short rate and to bond supply affect bond yields. Equation (22) implies that the effect of short-rate shocks on bond yields conforms to the EH: a unit short-rate shock raises the yield of the τ-period bond by $\partial y_t^{(\tau)}/\partial r_t = \tau^{-1}A_r^{(\tau)} = \tau^{-1}\sum_{j=0}^{\tau-1}(\rho_r)^j$—i.e., by the shock's effect on the average of the expected short rates over the next τ periods.

What is the intuition for this result? Holding bond expected returns $\mu_t^{(\tau)}$ fixed, a decline in the short rate would raise $\mu_t^{(\tau)} - r_t$ and, hence, arbitrageurs' demand for longer-term bonds. However, since the supply of bonds is fixed, the expected returns on bonds must adjust in equilibrium so that bond risk premia $\mu_t^{(\tau)} - r_t$ are not impacted by changes in the short rate. Since the convexity adjustment terms also do not depend on the short rate, the yield decomposition (10) implies that the effect of short-rate shocks on yields works through expected short rates only and conforms to the EH. By contrast, when the supply coming from preferred-habitat agents is elastic, as we will assume in Section 2.3, the effect of short-rate shocks on yields does not conform to the EH.

Equation (22) further implies that the sensitivity $A_r^{(\tau)}$ of bond prices to short-rate shocks rises with bond maturity τ, and the sensitivity $\tau^{-1}A_r^{(\tau)}$ of bond yields to short-rate shocks declines with maturity. Moreover, the effect of short-rate shocks is stronger when the short rate is more persistent (higher ρ_r).

Consider next shocks to bond supply. In line with the constant supply assumption in Section 2.2, we take supply shocks to be unanticipated and permanent. We represent a supply shock by a change $\Delta\zeta^{(\tau)}$ in the supply of τ-period bonds for $\tau = 1, 2, ..., T$. Since (22) implies that $A_r^{(\tau)}$ does not change, (12), (18) and (23) imply that the yield for maturity τ changes by

$$\Delta y_t^{(\tau)} = \frac{\Delta C^{(\tau)}}{\tau} = \frac{\sum_{j=1}^{\tau} A_r^{(j-1)}}{\tau}\underbrace{a\sigma_r^2 \sum\nolimits_{\tau=2}^{T} \Delta\zeta^{(\tau)} A_r^{(\tau-1)}}_{\Delta\lambda_r}. \tag{24}$$

The supply shock affects yields because it changes the dollar duration of arbitrageurs' portfolio with respect to the short rate. The dollar duration changes by $\sum_{\tau=2}^{T}\Delta\zeta^{(\tau)}A_r^{(\tau-1)}$, and the price of short-rate risk changes by $\Delta\lambda_r = a\sigma_r^2\sum_{\tau=2}^{T}\Delta\zeta^{(\tau)}A_r^{(\tau-1)}$. If portfolio dollar duration increases, then yields for all maturities rise—even for maturities for which supply decreases. This is illustrated in Figure 1 which shows the term structure response to an unanticipated supply shock in which the supply of 3-year bonds drops and the supply of 10-year bonds rises by an equal amount. Because the change in the supply of 10-year bonds raises portfolio dollar duration more than the change in the supply of 3-year bonds lowers it, yields for all maturities rise.

While changes in the supply of bonds of different maturities have different effects on yields, they all have the exact same *relative* effect across maturities, generating a term-structure response with the same shape. In other words, for any two maturities $\tau_2 > \tau_1$, the ratio $\Delta y_t^{(\tau_2)}/\Delta y_t^{(\tau_1)}$ is the same irrespective of the distribution of the supply shock $\Delta\zeta(\tau)$ across maturities τ. In that sense, supply effects are fully global. By contrast, with random supply shocks, assumed in Sections 2.4 and 3, supply effects become partially localized:

the shape of the term structure response depends on the distribution of the supply shock across maturities, and shifting that distribution towards longer maturities generates a term structure response that is relatively stronger for longer maturities.

Equation (24) also implies that the response of the term structure to supply shocks (of any maturity) increases with maturity: long-term bonds are more impacted than short-term bonds. Indeed, since $A_r^{(\tau)}$ is increasing in τ, so is $\tau^{-1}\sum_{j=1}^{\tau} A_r^{(j-1)}$. Intuitively, an increase in supply raises bond yields because it raises bond risk premia. Moreover, bond risk premia rise more for longer-maturity bonds because their prices are more sensitive to the short rate ($A_r^{(\tau)}$ increases in τ). Since supply changes are permanent, the average of the risk premia earned by a bond over its life increases more for longer-maturity bonds. However, supply shocks can instead have a hump-shaped effect on the term structure when they are random and mean-reverting as in Section 2.4, or when they are random and the supply coming from preferred-habitat agents is elastic as in Section 3.

An additional implication concerns the sign of bond risk premia and the average slope of the yield curve over time. In frictionless asset-pricing models—e.g., consumption-based models, bond risk premia are generally negative because short rates decline in "bad" economic times, making bonds a valuable hedge for diversified investors. This logic breaks down in our model of segmented markets, where the marginal investor in bonds is a specialized arbitrageur who holds a long position in bonds—i.e., $\lambda_r > 0$ if $\sum_{\tau=2}^{T}\zeta^{(\tau)}A_r^{(\tau-1)} > 0$. When short rates fall, specialized bond arbitrageurs earn large profits even though low short rates may correspond to bad times for the economy and for well-diversified investors. Thus, one appeal of the VV model is that it provides a natural economic explanation for why bond risk premia are typically positive and yield curves are upward-sloping on average.

2.3 Elastic Supply

In this section we assume that the supply coming from preferred-habitat agents responds elastically to bond prices. For simplicity, we assume that the supply of the τ-period bond depends only on the price of that bond and takes the form

$$S_t^{(\tau)} = \zeta^{(\tau)} + \eta^{(\tau)} p_t^{(\tau)} \tag{25}$$

for functions $\zeta^{(\tau)}$ and $\eta^{(\tau)}$ that depend only on τ. The function $\eta^{(\tau)}$ is the sensitivity of the supply of the τ-period bond to changes in its log price.

If $\eta^{(\tau)} > 0$, then bond supply is increasing in price. This could be because gross bond supply increases with price—e.g., because the government and corporations issue more debt when interest rates are low. It could also be because bond demand from other preferred-habitat agents decreases with price—e.g., investors substitute away from bonds and towards equities when interest rates are low, thereby reaching for yield across asset classes.

If $\eta^{(\tau)} < 0$, then bond supply is *decreasing* in price. This could be because gross bond supply decreases with price or bond demand from other preferred-habitat agents increases with price. Perhaps surprisingly, the literature has explored a number of mechanisms that can cause net bond supply to decrease with price. These include mortgage refinancing, asset-liability hedging by insurers and pensions, investors who extrapolate recent changes in short rates, and investors who reach for yield across the term structure (Hanson (2014), Hanson and Stein (2015), Malkhozov et al. (2016), Domanski et al. (2017), Hanson et al. (2021), Carboni and Ellison (2022)).

If there were no arbitrageurs in the model with elastic supply, then yields would have to adjust so that the net supply from preferred-habitat agents in (25) would equal to zero. Equilibrium yields would be given by

$$y_t^{(\tau)} = \frac{\zeta^{(\tau)}}{\eta^{(\tau)}\tau}. \tag{26}$$

As a result, yields would be constant over time and completely disconnected from changes in short rates, giving rise to severe failures of the EH. The market for each individual maturity would be completely segmented from that for other maturities, corresponding to an extreme form of the preferred-habitat view (Culbertson (1957), Modigliani and Sutch (1966)).

2.3.1 Deriving the Equilibrium

The derivation of the equilibrium begins as in Section 2.2.1 with constant supply. However, with elastic supply, $\lambda_{r,t}$ is not given by (18) but by

$$\lambda_{r,t} = a\sigma_r^2 \sum\nolimits_{\tau=2}^{T} [\zeta^{(\tau)} + \eta^{(\tau)} p_t^{(\tau)}] A_r^{(\tau-1)} = \lambda_{rr}(r_t - \overline{r}) + \lambda_r, \tag{27}$$

where

$$\lambda_{rr} \equiv -a\sigma_r^2 \sum\nolimits_{\tau=2}^{T} \eta^{(\tau)} A_r^{(\tau)} A_r^{(\tau-1)}, \tag{28}$$

$$\lambda_r \equiv a\sigma_r^2 \sum\nolimits_{\tau=2}^{T} \left[\zeta^{(\tau)} - \eta^{(\tau)} C^{(\tau)}\right] A_r^{(\tau-1)}. \tag{29}$$

Substituting $\mu_t^{(\tau)}$ from (13) and $\lambda_{r,t}$ from (27) into (16), we obtain a counterpart to (19)

$$\begin{aligned} &A_r^{(\tau)}(r_t - \overline{r}) - A_r^{(\tau-1)}\rho_r(r_t - \overline{r}) + C^{(\tau)} - C^{(\tau-1)} + \frac{\sigma_r^2}{2}[A_r^{(\tau-1)}]^2 - r_t \\ &= A_r^{(\tau-1)}\left[\lambda_{rr}(r_t - \overline{r}) + \lambda_r\right]. \end{aligned} \tag{30}$$

Identifying linear terms in $r_t - \overline{r}$, we find

$$A_r^{(\tau)} - A_r^{(\tau-1)}(\rho_r + \lambda_{rr}) - 1 = 0, \tag{31}$$

which is the counterpart of (20). Identifying constant terms, we find (21).

A complication when solving the difference equation (31) for $A_r^{(\tau)}$ is that the coefficient λ_{rr} depends on a sum involving $A_r^{(\tau)}$. This gives rise to a fixed-point problem, which we solve in two steps. First, we take $\rho_r^* \equiv \rho_r + \lambda_{rr}$ as given and solve (31) as a difference equation with constant coefficients. The solution is

$$A_r^{(\tau)} = \frac{1 - (\rho_r^*)^\tau}{1 - \rho_r^*}. \tag{32}$$

Second, we require that the solution (32) is consistent with the definition of λ_{rr} in (28). This gives rise to the following non-linear fixed-point condition in ρ_r^*:

$$\rho_r^* = \rho_r - a\sigma_r^2 \sum_{\tau=2}^{T} \eta^{(\tau)} \frac{1 - (\rho_r^*)^\tau}{1 - \rho_r^*} \frac{1 - (\rho_r^*)^{\tau-1}}{1 - \rho_r^*}. \tag{33}$$

ρ_r^* can be interpreted as the short rate's persistence under the risk-neutral measure.

Solving for $C^{(\tau)}$ involves solving a similar fixed-point problem. Taking λ_r as given, the solution of (21) is (23). Substituting (23) into (29) leads to a linear equation in λ_r whose solution is

$$\lambda_r = \frac{a\sigma_r^2 \sum_{\tau=2}^{T} \left[\zeta^{(\tau)} - \eta^{(\tau)}\tau\overline{r} + \frac{\sigma_r^2}{2}\eta^{(\tau)} \sum_{j=1}^{\tau} [A_r^{(j-1)}]^2\right] A_r^{(\tau-1)}}{1 + a\sigma_r^2 \sum_{\tau=2}^{T} \eta^{(\tau)} \left[\sum_{j=1}^{\tau} A_r^{(j-1)}\right] A_r^{(\tau-1)}}. \tag{34}$$

2.3.2 Results and Intuitions

When bond supply is elastic, short-rate shocks trigger an endogenous supply response from preferred-habitat agents, shifting bond risk premia. As a result, bond yields under- or over-react to short-rate shocks relative to the EH. Whether yields under- or over-react to the short rate depends on whether bond supply from preferred-habitat agents is increasing or decreasing in bond prices. When supply is upward-sloping ($\eta^{(\tau)} > 0$), bond risk premia are decreasing in the short rate and bond yields under-react to the short rate relative to the EH. The converse happens when supply is downward-sloping ($\eta^{(\tau)} < 0$).

To understand the intuition for under- and over-reaction, consider the case where bond supply is upward-sloping—i.e., where $\eta^{(\tau)} > 0$. Holding bond expected returns $\mu_t^{(\tau)}$ fixed, a decline in the short rate raises $\mu_t^{(\tau)} - r_t$ and, hence, arbitrageurs' demand for longer-term bonds. When supply is upward-sloping, arbitrageurs' buying pressure not only leads bond prices to rise but also leads arbitrageurs to buy more bonds from preferred-habitat agents. Because arbitrageurs hold more bonds when short rates fall, they become more exposed to future movements in short rates, so the price $\lambda_{r,t}$ of short-rate risk rises. Since $\lambda_{r,t}$ rises, bond risk premia rise and bond yields under-react to declines in the short rate relative to the EH baseline. Formally, when $\eta^{(\tau)} > 0$, equation (28) implies $\lambda_{rr} < 0$, so equation (27) implies that that $\lambda_{r,t}$ rises when r_t falls. As a result, when $\eta^{(\tau)} > 0$ and $\lambda_{rr} < 0$, we have $\rho_r^* \equiv \rho_r + \lambda_{rr} < \rho_r$—i.e., the persistence ρ_r^* of the short rate under the risk-neutral measure is smaller than the persistence ρ_r under the physical measure.

When $\eta^{(\tau)} > 0$, bond risk premia are positively related to the slope of the yield curve—i.e., to $y_t^{(\tau)} - r_t$ for $\tau \geq 2$. Indeed, a low short rate implies both a steeper than average yield curve and higher than average bond risk premia. As a result, a steep yield curve predicts higher future excess returns on bonds. The positive relationship between the slope of the yield curve and bond risk premia is one of the most widely documented empirical facts in the term-structure literature, starting with Fama and Bliss (1987) and Campbell and Shiller (1991).

Naturally, all of these results are reversed when $\eta^{(\tau)} < 0$. Specifically, when the supply from preferred-habitat agents is decreasing in bond prices, bond yields over-react to movements in short rates relative to the EH. As a result, versions of the VV model in which $\eta^{(\tau)} < 0$ have proven useful in understanding the "excess sensitivity" of long-term yields to changes in short rates that has been documented at higher frequencies. See, for example, Hanson (2014), Hanson and Stein (2015), Malkhozov et al. (2016), Domanski et al. (2017), Hanson et al. (2021), Carboni and Ellison (2022).

2.4 Random Supply

In this section we allow the supply coming from preferred-habitat agents to fluctuate randomly over time. We focus on the case where supply does not respond to bond prices. We analyze the case where supply is both random and responds elastically to bond prices in Section 3.

For simplicity, we assume that the random supply is driven by a single supply factor s_t. The supply of the τ-period bond is

$$S_t^{(\tau)} = \zeta^{(\tau)} + \theta^{(\tau)} s_t, \tag{35}$$

for functions $\zeta^{(\tau)}$ and $\theta^{(\tau)}$ that depend only on τ. We parameterize the function $\theta^{(\tau)}$ so that an increase in s_t raises the dollar duration of aggregate bond supply with respect to the short rate—i.e., $\sum_{\tau=2}^{T} \theta^{(\tau)} A_r^{(\tau-1)} > 0$. Dollar duration trivially increases in s_t when bond supply for all maturities increases in s_t—i.e., $\theta^{(\tau)} > 0$ for all $\tau \geq 2$.[6] We assume that the supply factor s_t follows the AR(1) process

$$s_{t+1} = \rho_s s_t + \sigma_s \varepsilon_{s,t+1}, \tag{36}$$

[6] A more general condition ensuring that an increase in s_t raises the dollar duration of aggregate bond supply with respect to the short rate is (i) $\sum_{\tau=2}^{T} \theta^{(\tau)} \geq 0$ and (ii) there exists $\tau^* \in [1, T-1]$ such that $\theta^{(\tau)} \leq 0$ for $\tau \leq \tau^*$ and $\theta^{(\tau)} > 0$ for $\tau > \tau^*$. Conditions (i) and (ii) imply $\sum_{\tau=2}^{T} \theta^{(\tau)} A_r^{(\tau-1)} > 0$ because the function $A_r^{(\tau)}$ is positive and increasing in τ.

where $\rho_s \in (0,1)$ and $\sigma_s > 0$ are constants, and $\varepsilon_{s,t+1}$ is a stochastic shock with $\mathbb{E}_t[\varepsilon_{s,t+1}] = 0$ and $\mathbb{V}\text{ar}_t[\varepsilon_{s,t+1}] = 1$. Setting the long-run mean of the s_t to zero is without loss of generality because we can redefine $\zeta(\tau)$. For simplicity, we also assume $\mathbb{C}\text{ov}_t[\varepsilon_{s,t+1}, \varepsilon_{r,t+1}] = 0$—i.e., supply shocks are uncorrelated with short-rate shocks—but this assumption can easily be relaxed. Greenwood and Vayanos (2014) study the continuous-time version of this model.

We look for an equilibrium where bond yields are affine functions of the short rate and the supply factor. In our conjectured equilibrium, the log price of the τ-period bond at time t takes the form

$$p_t^{(\tau)} = -\tau y_t^{(\tau)} = -\left[A_r^{(\tau)}(r_t - \overline{r}) + A_s^{(\tau)} s_t + C^{(\tau)}\right], \tag{37}$$

where the functions $A_r^{(\tau)}$, $A_s^{(\tau)}$ and $C^{(\tau)}$ depend only on maturity τ.

2.4.1 Deriving the Equilibrium

Substituting (37) into (7) and (8), and using the assumed AR(1) dynamics for the short rate and the supply factor, the expected and unexpected components of the return of the τ-period bond between times t and $t+1$ are

$$\begin{aligned} \mu_t^{(\tau)} &= A_r^{(\tau)}(r_t - \overline{r}) - A_r^{(\tau-1)}\rho_r(r_t - \overline{r}) + A_s^{(\tau)} s_t - A_s^{(\tau-1)}\rho_s s_t + C^{(\tau)} - C^{(\tau-1)} \\ &\quad + \frac{\sigma_r^2}{2}[A_r^{(\tau-1)}]^2 + \frac{\sigma_s^2}{2}[A_s^{(\tau-1)}]^2, \end{aligned} \tag{38}$$

$$r_{t+1}^{(\tau)} - \mu_t^{(\tau)} = -A_r^{(\tau-1)}\sigma_r\varepsilon_{r,t+1} - A_s^{(\tau-1)}\sigma_s\varepsilon_{s,t+1}, \tag{39}$$

respectively. Substituting (38) and (39) into the arbitrageurs' budget constraint (5) and using the assumption that supply shocks are uncorrelated with short-rate shocks, we can write the arbitrageurs' objective (6) as

$$\sum\nolimits_{\tau=2}^{T} X_t^{(\tau)}(\mu_t^{(\tau)} - r_t) - \frac{a\sigma_r^2}{2}\left(\sum\nolimits_{\tau=2}^{T} X_t^{(\tau)} A_r^{(\tau-1)}\right)^2 - \frac{a\sigma_s^2}{2}\left(\sum\nolimits_{\tau=2}^{T} X_t^{(\tau)} A_s^{(\tau-1)}\right)^2. \tag{40}$$

The arbitrageurs' first-order condition for $X_t^{(\tau)}$ is

$$\mu_t^{(\tau)} - r_t = A_r^{(\tau-1)}\lambda_{r,t} + A_s^{(\tau-1)}\lambda_{s,t}, \tag{41}$$

where the prices of factor risk are $\lambda_{f,t} \equiv a\sigma_f^2 \sum_{\tau=2}^{T} X_t^{(\tau)} A_f^{(\tau-1)}$ for $f = r, s$. Using the market-clearing condition $X_t^{(\tau)} = S_t^{(\tau)} = \zeta^{(\tau)} + \theta^{(\tau)} s_t$, the equilibrium prices of factor risk are

$$\lambda_{f,t} \equiv a\sigma_f^2 \sum\nolimits_{\tau=2}^{T}\left[\zeta^{(\tau)} + \theta^{(\tau)} s_t\right] A_f^{(\tau-1)} = \lambda_{fs} s_t + \lambda_f, \tag{42}$$

where

$$\lambda_{fs} \equiv a\sigma_f^2 \sum\nolimits_{\tau=2}^{T} \theta^{(\tau)} A_f^{(\tau-1)}, \tag{43}$$

$$\lambda_f \equiv a\sigma_f^2 \sum\nolimits_{\tau=2}^{T} \zeta^{(\tau)} A_f^{(\tau-1)}. \tag{44}$$

Substituting $\mu_t^{(\tau)}$ from (38) and $\lambda_{f,t}$ from (42) into (41), we find the equation

$$\begin{aligned} &A_r^{(\tau)}(r_t - \overline{r}) - A_r^{(\tau-1)}\rho_r(r_t - \overline{r}) + A_s^{(\tau)}s_t - A_s^{(\tau-1)}\rho_s s_t + C^{(\tau)} - C^{(\tau-1)} \\ &+ \frac{\sigma_r^2}{2}[A_r^{(\tau-1)}]^2 + \frac{\sigma_s^2}{2}[A_s^{(\tau-1)}]^2 - r_t = A_r^{(\tau-1)}[\lambda_{rs}s_t + \lambda_r] + A_s^{(\tau-1)}[\lambda_{ss}s_t + \lambda_s], \end{aligned} \tag{45}$$

which is affine in r_t and s_t. Identifying linear terms in $r_t - \overline{r}$, we find (20) from Section 2.2. Identifying linear terms in s_t, we find

$$A_s^{(\tau)} - A_s^{(\tau-1)}(\rho_s + \lambda_{ss}) - A_r^{(\tau-1)}\lambda_{rs} = 0. \tag{46}$$

Identifying constant terms, we find

$$C^{(\tau)} - C^{(\tau-1)} - \overline{r} - A_r^{(\tau-1)}\lambda_r - A_s^{(\tau-1)}\lambda_s + \frac{\sigma_r^2}{2}[A_r^{(\tau-1)}]^2 + \frac{\sigma_s^2}{2}[A_s^{(\tau-1)}]^2 = 0. \tag{47}$$

The solution for $A_r^{(\tau)}$ is (22) from Section 2.2. Solving for $A_s^{(\tau)}$ entails a similar fixed-point problem as in Section 2.3. Taking λ_{ss} as given, the solution of (46) is $A_s^{(1)} = 0$ and for $\tau > 2$

$$A_s^{(\tau)} = \lambda_{rs} \sum_{j=1}^{\tau-1} (\rho_s^*)^{(j-1)} A_r^{(\tau-j)} = \lambda_{rs} \frac{1}{\rho_s^* - \rho_r} \left(\frac{1 - (\rho_s^*)^{\tau}}{1 - \rho_s^*} - \frac{1 - \rho_r^{\tau}}{1 - \rho_r} \right) > 0, \tag{48}$$

where $\lambda_{rs} = a\sigma_r^2 \sum_{\tau=2}^{T} \theta^{(\tau)} A_r^{(\tau-1)} > 0$ and $\rho_s^* \equiv \rho_s + \lambda_{ss}$ satisfies the following fixed-point condition:

$$\rho_s^* = \rho_s + a\sigma_s^2 \lambda_{rs} \left[\sum\nolimits_{\tau=2}^{T} \theta^{(\tau)} \frac{1}{\rho_s^* - \rho_r} \left(\frac{1 - (\rho_s^*)^{\tau-1}}{1 - \rho_s^*} - \frac{1 - \rho_r^{\tau-1}}{1 - \rho_r} \right) \right]. \tag{49}$$

The parameter ρ_s^* can be interpreted as the persistence of the supply factor under the risk-neutral measure. Given our assumptions on $\theta^{(\tau)}$, any solution ρ_s^* of (49) satisfies $\rho_s^* > \rho_s$. The solution for $C^{(\tau)}$ is

$$C^{(\tau)} = \tau\overline{r} + \sum_{j=1}^{\tau} A_r^{(j-1)}\lambda_r + \sum_{j=1}^{\tau} A_s^{(j-1)}\lambda_s - \frac{\sigma_r^2}{2} \sum_{j=1}^{\tau} \left[A_r^{(j-1)}\right]^2 - \frac{\sigma_s^2}{2} \sum_{j=1}^{\tau} \left[A_s^{(j-1)}\right]^2. \tag{50}$$

2.4.2 Results and Intuitions

We first examine how shocks to bond supply affect bond yields and how the effect depends on the shocks' persistence. Recall from Section 2.2 that an unanticipated and permanent shock to the supply of bonds of any maturity has a larger effect on long-term yields than on short-term yields. Our results in this section imply that a mean-reverting supply shock can instead have a hump-shaped effect on the yield curve, which is largest for intermediate-maturity bonds. Formally, (37) and (48) imply that a unit shock to the supply factor s_t raises the yield for maturity τ by

$$\frac{\partial y_t^{(\tau)}}{\partial s_t} = \frac{A_s^{(\tau)}}{\tau} = \frac{\lambda_{rs}}{\tau} \frac{\frac{1-(\rho_s^*)^\tau}{1-\rho_s^*} - \frac{1-\rho_r^\tau}{1-\rho_r}}{\rho_s^* - \rho_r} = \frac{\lambda_{rs}}{\tau} \sum_{j=0}^{\tau-1} \frac{(\rho_s^*)^j - \rho_r^j}{\rho_s^* - \rho_r} \equiv \frac{\lambda_{rs}}{\tau} \sum_{j=0}^{\tau-1} \varphi^{(j)}. \tag{51}$$

When $\rho_s^* > 1$, the positive sequence $\varphi^{(j)}$ is increasing in j. Since $\partial y_t^{(\tau)}/\partial s_t$ is an average of an increasing sequence of numbers, it is increasing in τ. Therefore, the effect of a supply shock on the term structure increases with maturity. When instead $\rho_s^* < 1$, the positive sequence $\varphi^{(j)}$ is hump-shaped and converges to zero when τ goes to infinity. Therefore, $\partial y_t^{(\tau)}/\partial s_t$ is a hump-shaped function in τ, so supply shocks have a hump-shaped effect on the term structure. The condition $\rho_s^* < 1$ is met when ρ_s is small—i.e., the supply shock is transitory.

The intuition for the hump shape follows from the decomposition (10). A supply shock raises bond yields because it raises the average of the risk premia earned by a bond over its life. Risk premia earned at any given time after the shock rise more for longer-term bonds because their prices are more sensitive to the short rate and to supply.[7] However, risk premia are expected to decline over time following a supply shock because the shock is transitory, so risk premia are expected to remain elevated over a smaller fraction of the life of a longer-term bond. If supply shocks are sufficiently transitory, then their maximum impact will be on intermediate-term yields.

The existence of random supply shocks reinforces the positive relationship between risk premia and the slope of the yield curve that arises in Section 2.3 when supply is upward-sloping. An elevated supply from preferred-habitat agents implies that bond risk premia are higher than average because arbitrageurs must be induced to hold that supply. Holding short rates fixed, this means that long-term yields and the slope of the yield curve—i.e., $y_t^{(\tau)} - r_t$ for $\tau \geq 2$, are higher than average.

We next show that with random supply, the effects of supply are not fully global as in Section 2.2, but become instead partially localized. The relative effect of a supply shock on yields of different maturities depends on the distribution of the shock across maturities.

[7]Formally, equations (41) and (42) and the facts that $\lambda_{rs} > 0$ and $\lambda_{ss} > 0$ imply that the risk premium $\mu_t^{(\tau)} - r_t$ of τ-period bonds is increasing in s_t. Moreover, this effect becomes stronger for larger τ because $A_r^{(\tau-1)}$ and $A_s^{(\tau-1)}$ are increasing in τ.

In particular, the rise in long-term yields *relative* to the rise in short-term yields stemming from an increase in supply is greater when that increase concerns the supply of long-term bonds than when it concerns the supply of short-term bonds.

To explain the intuition for partial localization, consider an unanticipated and permanent change $\Delta\zeta^{(\tau)}$ in the supply of τ-period bonds for $\tau = 1, 2, ...T$. Since (22) and (48) imply $A_r^{(\tau)}$ and $A_s^{(\tau)}$ do not change, (37) and (50) imply that the yield for maturity τ changes by

$$\Delta y_t^{(\tau)} = \frac{\Delta C^{(\tau)}}{\tau} = \frac{\sum_{j=1}^{\tau} A_r^{(j-1)}}{\tau}\Delta\lambda_r + \frac{\sum_{j=1}^{\tau} A_s^{(j-1)}}{\tau}\Delta\lambda_s, \tag{52}$$

where $\Delta\lambda_f = a\sigma_f^2 \sum_{\tau=2}^{T} \Delta\zeta^{(\tau)} A_f^{(\tau-1)}$ for $f = r, s$ are the changes in the prices of short-rate and supply risk. Partial localization means that $\Delta y_t^{(\tau_2)}/\Delta y_t^{(\tau_1)}$ for any two maturities $\tau_2 > \tau_1$ is higher when the supply shock $\Delta\zeta^{(\tau)}$ originates at longer maturities. Key to partial localization is that $A_s^{(\tau)}/A_r^{(\tau)}$ is increasing in maturity τ—i.e., supply shocks are a more important driver of long-term yields whereas short-rate shocks are a more important driver of short-term yields. Indeed, when $A_s^{(\tau)}/A_r^{(\tau)}$ is increasing in τ, an increase $\Delta\zeta^{(\tau)}$ in the supply of long-term bonds results in a larger increase in the price of supply risk relative to the price of short-rate risk—i.e., a larger $\Delta\lambda_s/\Delta\lambda_r$, compared to a same-sized increase in the supply of short-term bonds. Since, in addition, $(\sum_{j=1}^{\tau_2} A_s^{(j-1)})/(\sum_{j=1}^{\tau_2} A_r^{(j-1)}) > (\sum_{j=1}^{\tau_1} A_s^{(j-1)})/(\sum_{j=1}^{\tau_1} A_r^{(j-1)})$ for $\tau_2 > \tau_1$, (52) implies that $\Delta y_t^{(\tau_2)}/\Delta y_t^{(\tau_1)}$ is higher when $\Delta\zeta^{(\tau)}$ originates at longer maturities.

3 The VV Model in Continuous Time

Our analysis of random supply shocks so far assumes that supply does not respond elastically to bond prices. In this section, we develop the continuous-time version of the VV model and use it to analyze the case where supply is both random and elastic. As we show below, in this case, even permanent shocks can have hump-shaped effects.

3.1 Model

Time t is continuous and infinite. At each time t, there is a continuum of zero-coupon, default-free bonds that have face value one and mature at time $t + \tau$ where $\tau \in (0, T]$. We denote the time-t price of the bond that matures at $t + \tau$ by $P_t^{(\tau)}$, and define the bond's yield to maturity $y_t^{(\tau)}$ and log price $p_t^{(\tau)}$ as in Section 2. Taking the unit of time to be years, we refer to the bond with maturity τ as the τ-year bond.

The short rate r_t is the limit of the yield $y_t^{(\tau)}$ as τ goes to zero. We take r_t as exogenous and assume that it follows the Orstein-Uhlenbeck process

$$dr_t = -\kappa_r(r_t - \bar{r})dt + \sigma_r dB_{r,t}, \tag{53}$$

where $\kappa_r > 0$, $\overline{r}$ and $\sigma_r > 0$ are constants, and $B_{r,t}$ is a Brownian motion. The process (53) is the continuous-time counterpart of (2). The persistence parameter ρ_r of the discrete-time AR(1) process (2) maps to $\exp(-\kappa_r \Delta t)$, where Δt is the time between discrete periods.

The instantaneous change in the price and log price of the τ-year bond at time t are denoted by $dP_t^{(\tau)}$ and $dp_t^{(\tau)}$, respectively. This differential accounts for the change in maturity—i.e., $dP_t^{(\tau)} = P_{t+dt}^{(\tau-dt)} - P_t^{(\tau)}$. Ito's lemma implies

$$\frac{dP_t^{(\tau)}}{P_t^{(\tau)}} = dp_t^{(\tau)} + \frac{1}{2}\mathbb{V}\text{ar}_t[dp_t^{(\tau)}]. \tag{54}$$

Equation (54) is the continuous-time counterpart of (3), but (54) holds exactly whereas (3) is only an approximation.

Arbitrageurs choose a bond portfolio to trade off the mean and variance of the instantaneous changes in their wealth W_t. The continuous-time counterpart of the budget constraint (5) is

$$dW_t = W_t r_t dt + \int_0^T X_t^{(\tau)} \left(\frac{dP_t^{(\tau)}}{P_t^{(\tau)}} - r_t \right) d\tau. \tag{55}$$

Arbitrageurs choose their bond holdings $\{X_t^{(\tau)}\}_{\tau\in(0,T]}$ to maximize

$$\mathbb{E}_t\left[dW_t\right] - \frac{a}{2}\mathbb{V}\text{ar}_t\left[dW_t\right]. \tag{56}$$

3.2 Elastic and Random Supply

We assume that the supply coming from preferred-habitat agents both fluctuates randomly over time and responds elastically to bond prices. The supply of the τ-year bond is

$$S_t^{(\tau)} = \zeta^{(\tau)} + \theta^{(\tau)} s_t + \eta^{(\tau)} p_t^{(\tau)} \tag{57}$$

for functions $\zeta^{(\tau)}$, $\theta^{(\tau)}$ and $\eta^{(\tau)}$ that depend only on τ. We assume that the supply factor s_t follows the Ornstein-Uhlenbeck (OU) process

$$ds_t = -\kappa_s s_t dt + \sigma_s dB_{s,t}, \tag{58}$$

where $\kappa_s > 0$ and $\sigma_s > 0$ are constants and $B_{s,t}$ is a Brownian motion. For simplicity, we assume that $B_{s,t}$ is independent of $B_{r,t}$—i.e., supply shocks are independent of short-rate shocks, but this assumption can easily be relaxed.

We look for an equilibrium where bond yields are affine functions of the short rate and the supply factor—i.e., where the log price of the τ-year bond at time t takes the form (37) in Section 2.4.

3.2.1 Deriving the Equilibrium

Applying Ito's lemma to $P_t^{(\tau)} = \exp(-[A_r^{(\tau)}(r_t - \overline{r}) + A_s^{(\tau)} s_t + C^{(\tau)}])$ and using the OU processes for the short rate and the supply factor, the instantaneous return of the τ-period bond at time t is

$$\frac{dP_t^{(\tau)}}{P_t^{(\tau)}} = \mu_t^{(\tau)} dt - A_r^{(\tau)} \sigma_r dB_{r,t} - A_s^{(\tau)} \sigma_s dB_{s,t}, \tag{59}$$

where

$$\mu_t^{(\tau)} \equiv \frac{dA_r^{(\tau)}}{d\tau}(r_t - \overline{r}) + \frac{dA_s^{(\tau)}}{d\tau} s_t + \frac{dC^{(\tau)}}{d\tau} + A_r^{(\tau)} \kappa_r (r_t - \overline{r}) + A_s^{(\tau)} \kappa_s s_t + \frac{\sigma_r^2}{2}\left[A_r^{(\tau)}\right]^2 + \frac{\sigma_s^2}{2}\left[A_s^{(\tau)}\right]^2 \tag{60}$$

denotes the instantaneous expected return. Equation (59) can also be derived from (54) together with (37), (53) and (58).

Substituting (59) into the arbitrageurs' budget constraint (55) and using the independence between short-rate and supply shocks, we can write the arbitrageurs' objective (56) as

$$\int_0^T X_t^{(\tau)} (\mu_t^{(\tau)} - r_t) d\tau - \frac{a}{2}\left[\sigma_r^2 \left(\int_0^T X_t^{(\tau)} A_r^{(\tau)} d\tau\right)^2 + \sigma_s^2 \left(\int_0^T X_t^{(\tau)} A_s^{(\tau)} d\tau\right)^2\right]. \tag{61}$$

Arbitrageurs maximize (61) over their bond holdings $\{X_t^{(\tau)}\}_{\tau \in (0,T]}$. Their first-order condition for $X_t^{(\tau)}$ is

$$\mu_t^{(\tau)} - r_t = A_r^{(\tau)} \lambda_{r,t} + A_s^{(\tau)} \lambda_{s,t}, \tag{62}$$

where the prices of factor risk are

$$\lambda_{f,t} \equiv a\sigma_f^2 \int_0^T X_t^{(\tau)} A_f^{(\tau)} d\tau \tag{63}$$

for $f = r, s$. Using the market-clearing condition $X_t^{(\tau)} = S_t^{(\tau)} = \zeta^{(\tau)} + \theta^{(\tau)} s_t + \eta^{(\tau)} p_t^{(\tau)}$, we find that the equilibrium prices of factor risk are

$$\lambda_{f,t} = a\sigma_f^2 \int_0^T [\zeta^{(\tau)} + \theta^{(\tau)} s_t + \eta^{(\tau)} p_t^{(\tau)}] A_f^{(\tau)} d\tau = \lambda_{fr}(r_t - \overline{r}) + \lambda_{fs} s_t + \lambda_f, \tag{64}$$

where

$$\lambda_{fr} \equiv -a\sigma_f^2 \int_0^T \eta^{(\tau)} A_r^{(\tau)} A_f^{(\tau)} d\tau, \tag{65}$$

$$\lambda_{fs} \equiv a\sigma_f^2 \int_0^T \left[\theta^{(\tau)} - \eta^{(\tau)} A_s^{(\tau)}\right] A_f^{(\tau)} d\tau, \tag{66}$$

$$\lambda_f \equiv a\sigma_f^2 \int_0^T \left[\zeta^{(\tau)} - \eta^{(\tau)} C^{(\tau)}\right] A_f^{(\tau)} d\tau. \tag{67}$$

Substituting $\mu_t^{(\tau)}$ from (60) and $\lambda_{f,t}$ from (64) into (62), we find an affine equation in r_t and s_t. Identifying linear terms in $r_t - \overline{r}$, linear terms in s_t and constant terms, we find

$$\frac{dA_r^{(\tau)}}{d\tau} + A_r^{(\tau)}(\kappa_r - \lambda_{rr}) - A_s^{(\tau)}\lambda_{sr} - 1 = 0, \tag{68}$$

$$\frac{dA_s^{(\tau)}}{d\tau} + A_s^{(\tau)}(\kappa_s - \lambda_{ss}) - A_r^{(\tau)}\lambda_{rs} = 0, \tag{69}$$

$$\frac{dC^{(\tau)}}{d\tau} - \overline{r} - A_r^{(\tau)}\lambda_r - A_s^{(\tau)}\lambda_s + \frac{\sigma_r^2}{2}\left[A_r^{(\tau)}\right]^2 + \frac{\sigma_s^2}{2}\left[A_s^{(\tau)}\right]^2 = 0, \tag{70}$$

respectively. Equations (68)-(70) constitute a system of three ordinary differential equations (ODEs). They must be solved with the terminal conditions $A_r^{(\tau)} = A_s^{(\tau)} = C^{(\tau)} = 0$.

The system of equations (68)-(70) reduces to solving one ODE at a time when supply is inelastic ($\eta^{(\tau)} = 0$) or non-random ($\sigma_s = 0$). In either case $\lambda_{sr} = 0$, so the term in $A_s^{(\tau)}$ drops from (68) and that ODE involves only $A_r^{(\tau)}$. Given its solution $A_r^{(\tau)}$, (69) can be solved for $A_s^{(\tau)}$, and given the solutions $A_r^{(\tau)}$ and $A_s^{(\tau)}$, (70) can be solved for $C^{(\tau)}$. Sections 2.2 to 2.4 use this recursive structure in discrete time. Specifically, when $\eta^{(\tau)} = 0$ or $\sigma_s = 0$, we have the following cases in continuous time, which map to Sections 2.2 to 2.4:

- When $\eta^{(\tau)} = 0$ and $\sigma_s = 0$, the solution for $A_r^{(\tau)}$ is

$$A_r^{(\tau)} = \underbrace{\int_0^\tau \exp(-\kappa_r j) dj}_{\frac{\partial}{\partial r_t}\left(\int_0^\tau \mathbb{E}_t[r_{t+j}]dj\right)} = \frac{1 - \exp(-\kappa_r \tau)}{\kappa_r}. \tag{71}$$

  The reaction of bond yields to changes in the short rate conforms to the EH. Equation (71) is the precise continuous-time analog of (22).

- When $\eta^{(\tau)} \neq 0$ and $\sigma_s = 0$, the solution for $A_r^{(\tau)}$ is

$$A_r^{(\tau)} = \frac{1 - \exp(-\kappa_r^* \tau)}{\kappa_r^*}, \tag{72}$$

  where $\kappa_r^* \equiv \kappa_r - \lambda_{rr}$, the short rate's mean reversion parameter under the risk-neutral measure, satisfies the fixed-point condition

$$\kappa_r^* = \kappa_r + a\sigma_r^2 \int_0^T \eta^{(\tau)} \left(\frac{1-\exp(-\kappa_r^*\tau)}{\kappa_r^*} \right)^2 d\tau. \tag{73}$$

Equation (73) implies $\kappa_r^* > \kappa_r$ when $\eta^{(\tau)} > 0$ (long-term yields under-react to the short rate relative to the EH) and $\kappa_r^* < \kappa_r$ when $\eta^{(\tau)} < 0$ (long-term yields over-react to the short rate relative to the EH). Equations (72) and (73) are the precise continuous-time analogs of (32) and (33).

- When $\eta^{(\tau)} = 0$ and $\sigma_s > 0$, the solution for $A_r^{(\tau)}$ is (71) and the solution for $A_s^{(\tau)}$ is

$$A_s^{(\tau)} = \lambda_{rs} \underbrace{\frac{1}{\kappa_r - \kappa_s^*} \left(\frac{1-\exp(-\kappa_s^*\tau)}{\kappa_s^*} \frac{1-\exp(-\kappa_r\tau)}{\kappa_r} \right)}_{\int_0^\tau \exp(-\kappa_s^* j) A_r^{(\tau-j)} dj}, \tag{74}$$

where $\lambda_{rs} = a\sigma_r^2 \int_0^T \theta^{(\tau)} A_r^{(\tau)} d\tau > 0$ and $\kappa_s^* \equiv \kappa_s - \lambda_{ss}$ satisfies the fixed-point condition

$$\kappa_s^* = \kappa_s - \lambda_{rs} a\sigma_s^2 \int_0^T \theta^{(\tau)} \frac{1}{\kappa_r - \kappa_s^*} \left(\frac{1-\exp(-\kappa_s^*\tau)}{\kappa_s^*} \frac{1-\exp(-\kappa_r\tau)}{\kappa_r} \right) d\tau. \tag{75}$$

The function $A_s^{(\tau)}$ is always increasing in τ while the function $\tau^{-1}A_s^{(\tau)}$ is increasing in τ when $\kappa_s^* < 0$ and is hump-shaped in τ when $\kappa_s^* > 0$. Equations (74) and (75) are the precise continuous-time analogs of (48) and (49).

When supply is both random and elastic, λ_{sr} is non-zero, and (68) and (69) must be solved as a system. Given the solution $A_r^{(\tau)}$ and $A_s^{(\tau)}$ of that system, (70) can be solved for $C^{(\tau)}$. To solve the system of (68) and (69), we write it in vector form

$$\begin{bmatrix} \frac{dA_r^{(\tau)}}{d\tau} \\ \frac{dA_s^{(\tau)}}{d\tau} \end{bmatrix} + \mathbf{M} \begin{bmatrix} A_r^{(\tau)} \\ A_s^{(\tau)} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \tag{76}$$

where

$$\mathbf{M} \equiv \begin{bmatrix} \kappa_r - \lambda_{rr} & -\lambda_{sr} \\ -\lambda_{rs} & \kappa_s - \lambda_{ss} \end{bmatrix} \equiv \begin{bmatrix} \kappa_r^* & -\lambda_{sr} \\ -\lambda_{rs} & \kappa_s^* \end{bmatrix}. \tag{77}$$

The system involves a fixed-point problem analogous to that in Sections 2.3 and 2.4 because the matrix $\mathbf{M}$ depends on integrals involving $A_r^{(\tau)}$ and $A_s^{(\tau)}$. We can solve the fixed-point problem in two steps, as in Sections 2.3 and 2.4. The first step is to take $\mathbf{M}$ as given and solve (68) and (69) as a linear ODE system with constant coefficients. The solutions are

$$A_r^{(\tau)} = \frac{1-\exp(-\nu_r\tau)}{\nu_r} - \frac{\nu_s-\kappa_s^*}{\nu_r-\nu_s}\left(\frac{1-\exp(-\nu_s\tau)}{\nu_s} - \frac{1-\exp(-\nu_r\tau)}{\nu_r}\right), \tag{78}$$

$$A_s^{(\tau)} = \lambda_{rs}\frac{1}{\nu_r-\nu_s}\left(\frac{1-\exp(-\nu_s\tau)}{\nu_s} - \frac{1-\exp(-\nu_r\tau)}{\nu_r}\right), \tag{79}$$

where

$$\nu_r = \frac{\kappa_r^*+\kappa_s^*+\sqrt{(\kappa_r^*-\kappa_s^*)^2+4\lambda_{sr}\lambda_{rs}}}{2} \text{ and } \nu_r = \frac{\kappa_r^*+\kappa_s^*-\sqrt{(\kappa_r^*-\kappa_s^*)^2+4\lambda_{sr}\lambda_{rs}}}{2} \tag{80}$$

are the eigenvalues of $\mathbf{M}$. The second step is to compute the integrals in $\mathbf{M}$ (i.e., to compute λ_{rr}, λ_{sr}, λ_{rs}, and λ_{ss}) given the solution $A_r^{(\tau)}$ and $A_s^{(\tau)}$, and require that they are consistent with the value of $\mathbf{M}$ used to compute $A_r^{(\tau)}$ and $A_s^{(\tau)}$. This problem amounts to solving a non-linear system of four scalar equations in the four elements of $\mathbf{M}$. When supply is inelastic or non-random, only one non-linear scalar equation needs to be solved—i.e., when $\lambda_{sr}=0$, $\nu_r=\kappa_r^*$ and $\nu_s=\kappa_s^*$.[8]

3.2.2 Results and Intuitions

We illustrate supply effects in a calibrated version of the model. We consider a negative shock $\Delta\zeta^{(\tau)}$ to bond supply $S_t^{(\tau)}$ that is unanticipated and permanent. All calibrated parameters except for the shock's persistence are as in Figure 3 of Vayanos and Vila (2021). The decrease in supply corresponds to QE purchases worth 12% of GDP.

Figure 2 plots the supply shock's effects on the term structure of yields. The left panel assumes that the shock is permanent. The right panel assumes that the shock reverts deterministically to zero at the rate $\kappa_\zeta = 0.15$, implying a half-life of 4.62 years. In each panel the red, green, light blue, blue, and black solid lines assume that the shock exclusively alters the supply of 2-, 5-, 10-, 20- and 30-year bonds, respectively. The black dashed line

[8] Two alternative approaches to solving this fixed-point problem have been proposed. One approach, developed in Hayashi (2018), uses the discrete-time version of the model and expresses the fixed-point problem as a system of $2T$ quadratic equations in the $2T$ unknowns $A_r^{(\tau)}$ and $A_s^{(\tau)}$ for $\tau = 1,..,T$. The unknowns are treated as functions of the arbitrageurs' risk aversion a, and an ODE in a is derived by differentiating implicitly the system with respect to a. The initial condition for the ODE is for the value $a=0$, for which the system can easily be solved in closed form.

Another approach, developed in Vayanos and Vila (2021), uses the continuous-time version of the model and assumes that T is infinite and the functions $\theta^{(\tau)}$ and $\eta^{(\tau)}$ are exponentials or linear combination of exponentials. The integrals in $\mathbf{M}$ then become Laplace transforms of the functions $A_r^{(\tau)}$ and $A_s^{(\tau)}$ and of their squares and product. Taking Laplace transforms of both sides of the ODE system (76), yields scalar equations in the Laplace transforms. These equations reduce to a non-linear system of four scalar equations. Relative to the approach of diagonalizing $\mathbf{M}$, the advantage of the Laplace transforms approach is that the fixed-point problem can be formulated and solved without needing to distinguish cases depending on whether the eigenvalues of $\mathbf{M}$ are real or complex.

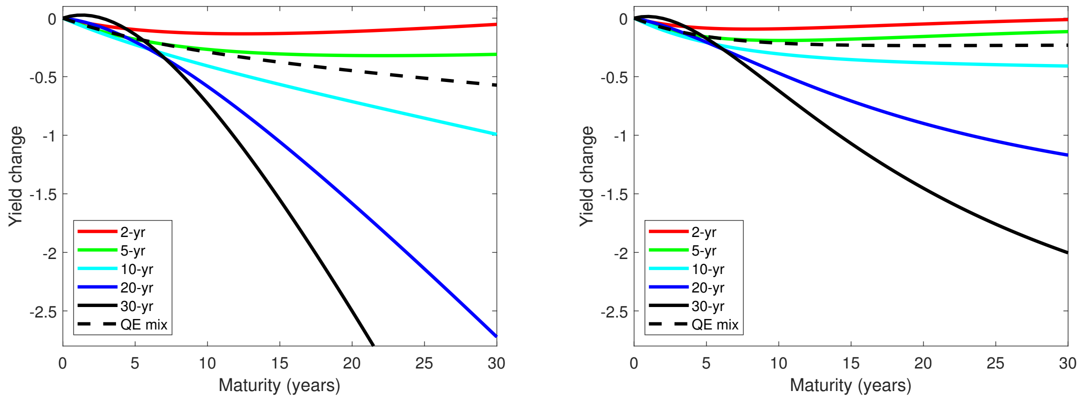

Figure 2: Effects of QE for the main calibration in VV.

assumes that the shock alters the supply of a basket of Treasury maturities whose composition matches the maturities purchased by the Fed during the first round of QE from March 2009 to March 2010—i.e., QE1—as reported in D'Amico and King (2013). Yield changes in the y-axis are in percent points—e.g., 2% is 200 basis points.

Figure 2 shows that supply effects exhibit some localization. Consistent with the dollar-duration intuition described in Section 2.2, a decrease in the supply of bonds with longer maturities generates a larger downward shift in the yield curve. For example, there is a larger downward shift when the negative supply shock is concentrated in 30-year bonds than when it is concentrated in 2-year bonds. That downward shift, however, is not larger for all maturities: yields on 1- to 3-year bonds are more sensitive to a decrease in supply of 2-year bonds than of 30-year bonds.

Figure 2 further shows that even permanent supply shocks can have hump-shaped effects on the yield curve. For example, a permanent decrease in the supply of 2-year bonds has its maximum effect on 10-year yields. Permanent shocks can have hump-shaped effects only when the supply coming from preferred-habitat agents is both random and elastic. Indeed, with a non-random or an inelastic supply, the effects are always increasing in maturity, as shown in Sections 2.2 and 2.4, respectively. The intuition for the hump shape is that because arbitrageurs sell bonds at the given maturity to accommodate the decrease in supply, they become more willing to buy bonds at other maturities to keep a similar risk profile. With elastic supply, the prices of those other bonds move only to the extent that arbitrageurs actually buy them. With two (or more) risk factors, bond returns are not perfectly correlated across maturities, so arbitrageurs are not willing to substitute as aggressively across maturities because doing so exposes them to additional risk. As a result, supply effects are partially concentrated around the maturities where the supply shocks land. Hump-shaped effects become more pronounced when supply is more elastic. They also become more pronounced when supply shocks are transitory, as can be seen by comparing the right panel of

Figure 2 to the left panel.

4 Perpetuities

For some applications, particularly in macroeconomics, the researcher may want to model the effects of supply and demand on long-term interest rates without modeling the entire yield curve. In this section, we replace the ladder of zero-coupon bonds with a single long-term bond, namely a coupon-bearing perpetual bond. Because coupons introduce nonlinearities, we linearize the perpetuity's return following Campbell and Shiller (1988). The resulting model is highly tractable and captures many of the insights in the previous sections. Simplified VV models of this type have been used in Greenwood et al. (2018), Hanson et al. (2021), Greenwood et al. (2023), and Hanson et al. (2023). This simple model is not a suitable framework for numerical calibration.

There are only two assets. The first is a 1-period bond. We refer to its yield, r_t, as the short rate and assume it follows the exogenous AR(1) process (2). The second asset is a perpetuity that has a face value of one and pays a coupon of $K > 0$ in each period. We denote the perpetuity's price and log price at time t by $P_t^{(L)}$ and $p_t^{(L)}$, respectively. We denote the perpetuity's return and log gross return between times t and $t+1$ by $R_{t+1}^{(L)} = (K + P_{t+1}^{(L)})/P_t^{(L)} - 1$ and $r_{t+1}^{(L)}$, respectively.

As in Section 2, we approximate the perpetuity's return between times t and $t+1$ by

$$R_{t+1}^{(L)} \approx r_{t+1}^{(L)} + \frac{1}{2}\mathbb{V}\mathrm{ar}_t[r_{t+1}^{(L)}]. \tag{81}$$

Following Campbell and Shiller (1988), we further approximate the perpetuity's log return as a linear function of its continuously compounded yield $y_t^{(L)}$. We carry out this approximation around the point where the perpetuity is trading at par—i.e., at a price equal to its face value of one. Specifically, we approximate the log return on the perpetuity from t to $t+1$ as

$$r_{t+1}^{(L)} \approx \frac{1}{1-\delta}y_t^{(L)} - \frac{\delta}{1-\delta}y_{t+1}^{(L)} = y_t^{(L)} - \frac{\delta}{1-\delta}(y_{t+1}^{(L)} - y_t^{(L)}), \tag{82}$$

where $\delta \equiv 1/(1+K) \in (0,1)$ is a constant of the linearization. The perpetuity's return is the sum of a "carry" component, y_t, which investors earn if yields do not change and a capital gain component, $-(\delta/(1-\delta))(y_{t+1} - y_t)$, due to changes in yields.[9]

Bond arbitrageurs choose portfolios consisting of 1-period bonds and perpetuities. They have mean-variance preferences over wealth 1-period ahead, with risk aversion a. Denoting by $X_t^{(L)}$ the market value of arbitrageurs' holdings of the perpetuity at time t, arbitrageur

[9]This log linear approximation dates back to Shiller, Campbell, and Schoenholtz (1983) and is discussed in Campbell (2018). To derive it, recall that the gross return on the perpetuity from t to $t+1$ is $1 + R_{t+1}^{(L)} = (K + P_{t+1}^{(L)})/P_t^{(L)}$. Letting $Y_t^{(L)} \equiv \exp(y_{t+1}^{(L)}) - 1$ denote the perpetuity's ordinary yield, the price of the

wealth evolves according to

$$W_{t+1} = (W_t - X_t^{(L)})(1 + R_t) + X_t^{(L)}(1 + R_{t+1}^{(L)}). \tag{83}$$

Using the approximation (3) for the return of the 1-period bond, and (81) for the return of the perpetuity, we can approximate the evolution of arbitrageurs' wealth by

$$W_{t+1} \approx \hat{W}_{t+1} \equiv W_t(1 + r_t) + X_t^{(L)}(r_{t+1}^{(L)} + \frac{1}{2}\mathbb{V}\mathrm{ar}_t[r_{t+1}^{(L)}] - r_t). \tag{84}$$

Arbitrageurs choose their holdings of the perpetuity $X_t^{(L)}$ to maximize (6). Defining $\mu_t^{(L)} \equiv \mathbb{E}_t[r_{t+1}^{(L)}] + \frac{1}{2}\mathbb{V}\mathrm{ar}_t[r_{t+1}^{(L)}]$, this is equivalent to maximizing

$$(\mu_t^{(L)} - r_t)X_t^L - \frac{a}{2}\mathbb{V}\mathrm{ar}_t[r_{t+1}^{(L)}](X_t^{(L)})^2. \tag{85}$$

The arbitrageurs' first-order condition for $X_t^{(L)}$ is

$$\mu_t^{(L)} - r_t = a\mathbb{V}\mathrm{ar}_t[r_{t+1}^{(L)}]X_t^{(L)}. \tag{86}$$

We assume that the supply of the perpetuity coming from preferred-habitat agents is

$$S_t^{(L)} = \zeta + s_t - \eta y_t^{(L)}, \tag{87}$$

where ζ and η are constants, and s_t follows the AR(1) process in equation (35).

We look for an equilibrium where the yield of the perpetuity is an affine function of the short rate r_t and the supply factor s_t. The yield takes the form

$$y_t^{(L)} = B_r(r_t - \overline{r}) + B_s s_t + B, \tag{88}$$

---

perpetuity is $P_t^{(L)} = K/Y_t^{(L)} = K/(\exp(y_t^{(L)}) - 1)$. Thus, the log return on the perpetuity from t to $t+1$ is

$$r_{t+1}^{(L)} = \log\left(\frac{K + P_{t+1}^{(L)}}{P_t^{(L)}}\right) = \log\left(K + \frac{K}{\exp(y_{t+1}^{(L)}) - 1}\right) - \log\left(\frac{K}{\exp(y_t^{(L)}) - 1}\right).$$

Linearizing this equation about the point where the perpetuity is trading at par at times t and $t+1$—i.e., about the point where $y_{t+1}^{(L)} = y_t^{(L)} = \log(1 + K)$, we find

$$r_{t+1}^{(L)} \approx y_{t+1}^{(L)} - \frac{1+K}{K}\left(y_{t+1}^{(L)} - \log(1+K)\right) + \frac{1+K}{K}\left(y_t^{(L)} - \log(1+K)\right) = \frac{1+K}{K}y_t^{(L)} - \frac{1}{K}y_{t+1}^{(L)},$$

which coincides with (82). Equation (82) can be generalized to allow the perpetuity to be self-amortizing, enabling a modeler to control the value of δ. In each period, a self-amoritizing perpetuity pays the coupon K and a fraction $1 - \lambda$ of its face value, and leaves the owner with λ units of the same perpetuity, where $\lambda \in [0, 1]$. (The standard perpetuity corresponds to $\lambda = 1$.) The linearization constant for a self-amortizing perpetuity is $\delta = \lambda/(1 + K)$. See Greenwood et al. (2023).

for constants B_r, B_s and B. This conjecture and the approximation (82) imply that the conditional variance of 1-period perpetuity returns is constant over time and given by

$$\mathbb{V}\mathrm{ar}_t[r_{t+1}^{(L)}] = \mathbb{V}\mathrm{ar}^{(L)} \equiv \left(\frac{\delta}{1-\delta}\right)^2 \left(B_r^2\sigma_r^2 + B_s^2\sigma_s^2\right) > 0. \tag{89}$$

To solve the model, we substitute $X_t^{(L)}$ from the market-clearing condition $X_t^{(L)} = S_t^{(L)}$ into the arbitrageurs' first-order condition (86). Using the definition of $\mu_t^{(L)}$, the approximations (81) and (82), the conjectured form of yields (88) and the AR(1) dynamics for the short rate and the supply factor, we find the equation

$$\begin{aligned} &\overbrace{\frac{1-\delta\rho_r}{1-\delta}B_r(r_t-\overline{r}) + \frac{1-\delta\rho_s}{1-\delta}B_s s_t + B + \frac{1}{2}\mathbb{V}\mathrm{ar}^{(L)} - r_t}^{\mu_t^{(L)}-r_t} \\ &= a\mathbb{V}\mathrm{ar}^{(L)}\left(\zeta + s_t - \eta\left[B_r(r_t-\overline{r}) + B_s s_t + B\right]\right), \end{aligned} \tag{90}$$

which is affine in $r_t - \overline{r}$ and s_t. Identifying linear terms in $r_t - \overline{r}$, linear terms in s_t and constant terms, we find

$$B_r = \frac{\frac{1-\delta}{1-\delta\rho_r}}{1 + \frac{1-\delta}{1-\delta\rho_r}\eta a\mathbb{V}\mathrm{ar}^{(L)}} > 0, \tag{91}$$

$$B_s = \frac{\frac{1-\delta}{1-\delta\rho_s}a\mathbb{V}\mathrm{ar}^{(L)}}{1 + \frac{1-\delta}{1-\delta\rho_s}\eta a\mathbb{V}\mathrm{ar}^{(L)}} > 0, \tag{92}$$

$$B = \overline{r} + \frac{a\mathbb{V}\mathrm{ar}^{(L)}(\zeta - \eta(\overline{r} - \frac{1}{2}\mathbb{V}\mathrm{ar}^{(L)}))}{1 + \eta a\mathbb{V}\mathrm{ar}^{(L)}} - \frac{1}{2}\mathbb{V}\mathrm{ar}^{(L)}, \tag{93}$$

respectively. Equations (91)-(93) are the counterparts of the difference and differential equations for $A_r^{(\tau)}$, $A_s^{(\tau)}$ and $C^{(\tau)}$ derived in Sections 2 and 3. They are scalar equations, which makes the perpetuity model simpler to solve.

Because B_r and B_s depend on $\mathbb{V}\mathrm{ar}^{(L)}$, and $\mathbb{V}\mathrm{ar}^{(L)}$ depends on B_r and B_s, a fixed-point problem analogous to that in Sections 2 and 3 arises. We can reduce the fixed-point problem to a scalar equation in $\mathbb{V}\mathrm{ar}^{(L)}$, as can be seen by combining (89), (91), and (92):

$$\mathbb{V}\mathrm{ar}^{(L)} = \left(\frac{\delta}{1-\delta}\right)^2 \left(\left(\frac{\frac{1-\delta}{1-\delta\rho_r}}{1 + \frac{1-\delta}{1-\delta\rho_r}\eta a\mathbb{V}\mathrm{ar}^{(L)}}\right)^2 \sigma_r^2 + \left(\frac{\frac{1-\delta}{1-\delta\rho_s}a\mathbb{V}\mathrm{ar}^{(L)}}{1 + \frac{1-\delta}{1-\delta\rho_s}\eta a\mathbb{V}\mathrm{ar}^{(L)}}\right)^2 \sigma_s^2\right).$$

When supply is constant and inelastic ($\sigma_s^2 = \eta = 0$), the fixed-point problem is degenerate and there is a unique equilibrium. When instead $\sigma_s^2 > 0$ or $\eta \neq 0$, the fixed-point problem is non-degenerate because the risk of holding perpetuities depends on how their prices react to shocks, and vice-versa. Furthermore, when $\sigma_s^2 > 0$ or $\eta < 0$, the fixed-point problem can have multiple solutions. Researchers typically focus on the unique equilibrium that is stable and does not explode in the limit where σ_s^2 and η go to zero.

Many—but not all—of the qualitative implications of the VV model can be illustrated in the perpetuity model. For example, as in Section 2.3 above, the sign of η determines how changes in the short rate impact bond risk premia and, thus, whether long-term yields under- or over-react to the short rate relative to the EH. When $\eta = 0$, the reaction of long-term yields to changes in the short rate conforms to the EH—i.e., $\partial(\mu_t^{(L)} - r_t)/\partial r_t = 0$ and bond risk premia do not depend on short rates, so

$$B_r = \frac{\partial}{\partial r_t}(1-\delta)\sum\nolimits_{j=0}^{\infty} \delta^j \mathbb{E}_t[r_{t+j}] = \frac{1-\delta}{1-\delta\rho_r}.$$

When $\eta > 0$, long-term yields to under-react to movements in the short rate relative to the EH because $\partial(\mu_t^{(L)} - r_t)/\partial r_t < 0$, and $B_r < (1-\delta)/(1-\delta\rho_r)$. Long-term yields instead over-react to movements in the short rate relative to the EH when $\eta < 0$. Similarly, as in Section 2.4, when $\eta = 0$, the impact of random supply shocks on long-term yields is greater when ρ_s is larger and supply shocks are more persistent.

5 Empirical applications of the VV framework

Empirical applications of the VV framework have sought to either isolate shocks to the supply or demand for long-term bonds, or alternately have sought to construct empirical proxies for arbitrageurs' holdings of long-term bonds and risk aversion. Several papers have further argued that some agents in bond markets trade in a "rate-amplifying" fashion in the short-run, causing long rates to over-react to movements in the short rate relative to the EH at high frequencies (Hanson et al. (2021)). Our goal in this section is not to provide an exhaustive treatment of the empirical literature on supply-and-demand effects in the term structure. Instead, we seek to connect some of the main findings in this growing empirical literature to concepts emphasized in the VV model.

5.1 Variation in bond supply

Garbade and Rutherford (2007) and Greenwood and Vayanos (2010) study the effects of a significant shift in the supply of long-term bonds: the U.S. Treasury's 2000−2001 debt buyback program. In early 2000, the U.S. Treasury, which at the time was running budget surpluses, announced that it would begin repurchasing long-term bonds. Between March 2000 and December 2001, the Treasury repurchased a significant fraction of outstanding debt maturing in more than 10 years. Consistent with the idea of preferred-habitat, Garbade and Rutherford (2007) and Greenwood and Vayanos (2010) find that the announcement of these debt repurchases was associated with a large drop in 30-year yields relative to 10-year yields.

Greenwood and Vayanos (2014) conduct a more systematic historical examination of how the supply and maturity structure of government debt impacts long-term yields in the U.S. To do so, they compute the maturity-weighted debt-to-GDP ratio, which is roughly the product of the debt-to-GDP ratio times the weighted-average-maturity of government debt, analogous to the dollar duration concept in (18). Greenwood and Vayanos (2014) then regress long-term yields and the future excess returns on long-term bonds on this duration supply variable, controlling for the level of short rates. They find that duration supply is positively related to current yield spreads relative to the short rate and predicts future excess returns on long-term bonds.

Many papers have studied the impact of central bank QE policies on bond yields. Gagnon et al. (2011) report that the combined impact of U.S. QE announcement events between 2008 and 2010 was to reduce 10-year U.S. Treasury yields by 91 basis points and and 10-year agency yields by 156 basis points. Similar effects have been documented in other countries, including the U.K. and the Eurozone (Joyce et al. (2011), Fratzscher et al. (2018)). Although most studies focus on the overall steepening of the yield curve, D'Amico and King (2013) document local effects in the spirit of the version of the model discussed in Section 3. Bhattarai and Neely (2022) provide a comprehensive review of the empirical literature on unconventional monetary policy.

Moving beyond government bonds, Hanson (2014) and Malkhozov et al. (2016) argue that mortgage refinancing waves in the U.S. are associated with significant shifts in the effective supply of long-term bonds that must be held by arbitrageurs. Most home mortgages in the U.S. are 30-year fixed-rate loans with a no-penalty prepayment option. When long-term interest rates decline, more households are expected to refinance their mortgages in the near term, so the ensuing refinancing waves cause the effective maturity of outstanding mortgages to shrink. As a result, mortgage refinancing waves lead to significant shifts in the total dollar quantity of interest rate risk that U.S. fixed-income investors must bear in equilibrium. Both Hanson (2014) and Malkhozov et al. (2016) build VV-style models that feature endogenous mortgage refinancing waves. In reduced form, both models are similar to assuming that $\eta < 0$—i.e., that the net supply of long-term bonds is downward-sloping and thus increasing in the level of long-term bond yields. Empirically, these papers show that the duration of outstanding mortgage-backed securities is a strong and positive predictor of future declines in long-term yields and, hence, the future excess returns on long-term bonds. Moreover, while the impact of mortgage duration on expected excess returns is increasing in bond maturity, increases in mortgage duration are short-lived in nature and have a hump-shaped effect on the yield curve, with their maximum impact being felt on intermediate maturities.

5.2 Variation in bond demand

Greenwood and Vayanos (2010) study a large shock to the demand for long-term bonds stemming from the 2004 reforms to pension regulation in the U.K. These reforms increased the incentives of pension funds to hold long-term U.K. government bonds, especially long-term inflation-protected securities. Consistent with the VV framework, this increase in demand for long-term bonds was associated with an inversion of the yield curve and an especially large effect on the long end: long-term U.K. yields (e.g., 30-year yields) declined significantly relative to intermediate-term yields (10-year yields). Because of these 'local' effects, the empirical results are best understood in the context of the version of the model with elastic and random supply.

Greenwood and Vissing-Jorgensen (2018) conduct a more systematic cross-country examination of how the demand for long-term bonds from pension funds and insurance companies affects the long end of the yield curve. Specifically, using data from 26 countries, they find that the spread between 30-year and 10-year government bond yields is negatively related to the ratio of pension and life insurance assets to GDP. This finding suggests that preferred-habitat demand from pensions and insurers for long-term bonds plays an important role in shaping the long end of the yield curve. Exploiting regulatory changes in several European countries from 2008 and 2013, Greenwood and Vissing-Jorgensen (2018) provide event-study evidence on the effect, further supporting the idea that pension and insurance demand impacted the long end of the yield curve. In particular, they argue that pension and insurance demand is partially driven by hedging linked to the regulatory discount curves. When regulators reduce the dependence of the regulatory discount curve on a particular security, pension and insurance demand for the security falls and its yield increases. Similar to Greenwood and Vayanos (2010), Greenwood and Vissing-Jorgensen (2018) provide several examples whereby demand shocks have local effects. In this way, they are best interpreted as an illustration of the VV model with elastic and random supply.

Ray et al. (2023) use news about demand revealed during Treasury auctions to measure how shocks transmit through the term structure. Their empirical strategy is based on the idea that all supply information is known before the close of each auction. Therefore, they argue, the release of the auction results reveals unexpected shifts in demand. By utilizing high frequency changes in Treasury yields around the close of Treasury auctions, they document that demand shocks are large and persistent, with effects on yields that last for many weeks following the auction. They extrapolate their findings to understand the impact of QE.

Domanski et al. (2017) point to a rate-amplification (or convexity-driven) mechanism stemming from the desire of pensions and insurers to dynamically match the duration of their assets and liabilities. Specifically, since the convexity of their liabilities exceeds that of their assets, pensions and insurers tend to purchase additional long-term bonds when long-term interest rates fall, to dynamically manage their interest-rate exposure. Empirically, Domanski et al. (2017) examine the portfolio rebalancing behavior of German insurers and argue that their convexity-dirven portfolio rebalancing contributed to the large decline in long-term interest rates in the Eurozone in late 2014. As above, these findings could be explained in a VV-style model where $\eta < 0$.

Hanson and Stein (2015) provide further empirical evidence for a negative η in the VV model. They show that long-term real yields on U.S. Treasury bonds are surprisingly sensitive to unexpected movements in short-term nominal interest rates driven by central bank policy announcements. They also propose a different institutional explanation for why η might be negative. The idea is that some investors care about the current yield on their portfolio as opposed to the portfolio's expected return. When short-term yields decline, EH logic implies that long-term yields decline by less, causing the term structure to steepen. This boosts the demand for long-term bonds from yield-seeking investors—i.e., these investors "reach for yield" and buy more long-term bonds when rates drop, pushing down the risk premium component of long-term yields.

5.3 Variation in arbitrageur risk tolerance

Although the VV model specifies the first-order condition of arbitrageurs, it is silent on who the relevant arbitrageurs are. Haddad and Sraer (2020) suggest that banks are marginal investors in fixed-income markets (through lending decisions, willingness to hold Treasury and Agency bonds, and so on), and that, therefore, the amount of interest-rate risk that banks are holding should determine bond risk premia and future excess returns. Haddad and Sraer (2020) measure bond risk premia through banks' average "income gap" computed as bank assets that reprice within one year, minus liabilities that mature or reprice within one year. The lower the income gap is, the higher is the interest-rate risk that banks must bear. Haddad and Sraer (2020) show that the income gap forecasts excess bond returns over the next four quarters, with increasing magnitudes for bonds of longer maturities.

6 Extensions to models of default-free government bonds

In this section, we discuss extensions to the baseline VV model that have been developed in the context of default-free bonds. We sketch the key formal elements of each extension and its relationship to the analysis in Sections 2 and 3.

6.1 Multiple supply factors

In Sections 2 and 3, random supply arises from a single supply factor. In practice, however, a variety of supply factors may operate at different maturities. For example, a factor reflecting the demand of mutual-fund managers may shift the (net) supply of short-term bonds, and a factor reflecting the demand of pension funds and insurance companies may shift the (net) supply of long-term bonds.

We can generalize the model for $K > 1$ supply factors by assuming that the supply of τ-period bonds at time t is

$$S_t^{(\tau)} = \zeta^{(\tau)} + \sum\nolimits_{k=1}^{K} \theta_k^{(\tau)} s_{k,t}, \tag{94}$$

for functions $\zeta^{(\tau)}$ and $\{\theta_k^{(\tau)}\}_{k=1}^K$ that depend only on τ, and for supply factors $\{s_{k,t}\}_{k=1}^K$. Each supply factor follows an AR(1) process of the form (36) in discrete time, or an OU process of the form (58) in continuous time. In either case, equilibrium bond yields are affine functions of the short rate and the K supply factors.

The model can be solved following the steps outlined in Sections 2 and 3. First, we compute $\mu_t^{(\tau)}$ using the conjectured bond yields, short-rate dynamics and supply-factor dynamics. Second, we derive the arbitrageurs' first-order condition for $X_t^{(\tau)}$. For example, assuming that time is discrete and the short rate and the K supply factors are all mutually independent, we find

$$\mu_t^{(\tau)} - r_t = A_r^{(\tau-1)} \lambda_{r,t} + \sum\nolimits_{k=1}^{K} A_{s_k}^{(\tau-1)} \lambda_{s_k,t}, \tag{95}$$

where the prices of risk are $\lambda_{f,t} \equiv a\sigma_f^2 \sum_{\tau=2}^{T} X_t^{(\tau)} A_f^{(\tau-1)}$ for $f = r, \{s_k\}_{k=1}^K$, and $A_{s_k}^{(\tau)}$ is the sensitivity of bond prices to supply factor s_k. Third, we determine the equilibrium prices of risk by imposing market clearing, $X_t^{(\tau)} = S_t^{(\tau)}$, where $S_t^{(\tau)}$ is given by (94). This yields an affine equation in the short rate and the supply factors that we use to construct a system of difference equations in discrete time or a system of ODEs in continuous time. As in Sections 2 and 3, that system can be solved by first treating it as system with constant coefficients and then requiring that (some of) these coefficients satisfy a fixed-point condition.

With multiple supply factors, the effects of supply shocks are more localized than with a single supply factor. Intuitively, supply effects become more localized because the arbitrage trades required to spread and smooth their effects globally are risky for arbitrageurs and hence are not undertaken aggressively in equilibrium. A model with multiple supply factors, including a supply factor that acts primarily on longer-term bonds, helps further understand the finding from Greenwood and Vayanos (2010) and Greenwood and Vissing-Jorgensen (2018) that shifts in the demand for long-term bonds from pensions and insurers can drive down 30-year yields relative to 10-year yields.

6.2 Forward guidance

In Sections 2 and 3, expectations about future short rates and bond supply change only when the current values of these variables change. This is because the variables follow AR(1) or OU processes. In practice, however, expectations about future short rates and bond supply can change even when the current values of these variables do not change. For example, central banks engage in forward guidance, conveying their intentions about the future path of short rates, without necessarily moving current short rates. Likewise, central banks have typically conducted QE by announcing a future path of bond purchases rather than by changing bond supply upon announcement.

We can introduce shocks to expected future short rates and bond supply by allowing the short rate and the supply factor to revert to a time-varying mean. For example, in discrete time, we can assume that factor f for $f = r, s$ evolves according to

$$f_{t+1} = \overline{f}_t + \rho_f (f_t - \overline{f}_t) + \varepsilon_{f,t+1}, \tag{96}$$

where the time-varying mean $\overline{f}_t$ evolves according to

$$\overline{f}_{t+1} = \overline{f} + \rho_{\overline{f}} (\overline{f}_t - \overline{f}) + \varepsilon_{\overline{f},t+1}. \tag{97}$$

A natural assumption under this specification is that shocks to a factor's time-varying mean are more persistent than shocks to the factor's current value ($0 < \rho_f < \rho_{\overline{f}} < 1$).

Greenwood et al. (2016) model forward guidance on short rates and bond supply using the specification (96) and (97). Equilibrium bond yields are affine functions of the short rate, the supply factor, and their time-varying means.

Shocks to expected future short rates have a hump-shaped effect on the yield curve because investors expect the shocks to be temporary. The location of the hump depends on the shocks' persistence. For example and in line with the EH, if investors expect short rates to rise and reach their maximum value in five years, then the hump will be located at a longer maturity than if investors expect short rates to reach their maximum in three years.

Shocks to expected future supply have a hump-shaped or an increasing effect on the yield curve. Moreover, the hump is located at a longer maturity compared to shocks to expected future short rates with the same persistence ($\rho_{\overline{r}} = \rho_{\overline{s}}$). This is because shocks to expected future supply affect bond yields through expected future bond risk premia, and risk premia primarily affect longer-term bonds.

6.3 Effective lower bound

Under the AR(1) and OU processes in Sections 2 and 3, the short rate can take arbitrarily low negative values. King (2019) incorporates an effective lower bound (ELB) on the short rate into the VV framework. The short rate is assumed to be $r_t = \max[b, \hat{r}_t]$, where $\hat{r}_t$ is a "shadow" short rate that follows an OU process and b is a lower bound, which can be set to zero.

In the presence of an ELB, equilibrium bond yields become nonlinear functions of the short rate and bond supply, and the model must be solved numerically. The nonlinearities give rise to three key insights concerning behavior close to the ELB. First, forward guidance that future short rates will be kept low—i.e., news that $\hat{r}_t$ has fallen—lowers uncertainty about future short rates and bond yields. This is because future short rates will be close to the ELB, and thus will exhibit little variation. Second, forward guidance that future short rates will be low reduces bond risk premia, holding bond supply fixed, because of the lower uncertainty. Third, QE purchases have a smaller effect on bond yields close to the ELB. This is because short-rate volatility is low close to the ELB, so arbitrageurs do not bear as much short-rate risk per unit of supply as they would have to bear away from the ELB.

6.4 Slow-moving capital

Greenwood et al. (2018) and Hanson et al. (2021) add a slow-moving arbitrage response to the VV model. Specifically, following Duffie (2010), these papers assume that some fraction of arbitrageurs are slow-moving and only periodically rebalance their portfolios. This assumption implies that arbitrageurs' short-run demand curve is less price-elastic than their long-run demand curve. Hence, the short-run effect of supply shocks on bond yields exceeds the shocks' long-run effect.

These models generate a distinction between "stock" and "flow" effects. In the baseline VV model, there is no distinction because (i) flows are simply changes in stocks and (ii) the slope of arbitrageurs' demand curve is not horizon-dependent. As a result, the current stock of supply fully characterizes bond risk premia. However, once the slopes of arbitrageurs' short- and long-run demand curves differ, the path of supply shocks matters: a jump in supply has a larger (temporary) effect on bond yields than a gradual increase in supply of the same cumulative magnitude.

6.5 Wealth effects and balance-sheet constraints

The risk aversion of arbitrageurs in the baseline VV model can be viewed as a reduced form that arises from contracting frictions between specialized bond arbitrageurs and the highly diversified investors who ultimately provide these arbitrageurs with capital. Specifically, due to moral hazard problems, specialized bond arbitrageurs can only raise capital from more diversified investors if arbitrageurs' compensation is tightly linked to the returns on their bond portfolios. Arbitrageur risk aversion can also be viewed as a reduced form for the constraints that regulators impose on arbitrageurs. Researchers have extended VV in two directions to account more explicitly for how contracting frictions and regulatory constraints affect arbitrageurs' demand and the term structure.

Kekre et al. (2023) assume that the risk aversion of arbitrageurs is inversely proportional to their wealth, instead of being a constant as in VV. Risk aversion that decreases in wealth can be a better reduced form for contracting frictions than constant risk aversion because it captures the notion that arbitrageurs become more constrained following losses. Risk aversion that is inversely proportional to wealth also results from intertemporal optimization of arbitrageurs who have log utility over their consumption.

With wealth-dependent risk aversion, the model can be solved analytically only in discrete time and with 1- and 2-period bonds. Otherwise, arbitrageur wealth becomes a state variable and only numerical solutions are possible. A key insight is that when arbitrageurs hold bond portfolios with positive duration, an unexpected drop in the short rate revalues wealth in their favor, lowering bond risk premia, and causing long rates to over-react relative to the EH. Wealth effects thus provide an alternative explanation to a downward-sloping supply curve (Section 2.3) for the over-reaction of long rates to the short rate.

He et al. (2022), Hanson et al. (2023) and Greenwood et al. (2023) maintain the VV assumption that arbitrageur risk aversion is constant but introduce balance-sheet costs, which further limit arbitrageurs' willingness to absorb supply shocks. These costs often stem from regulations and, importantly, can apply even to completely riskless trades. For example, since the 2008 Global Financial Crisis, large dealer banks have been subject to a non-risk-based equity capital requirement called the Supplemental Leverage Ratio (SLR). When the SLR binds, banks are required to finance even riskless trades using some amount of equity capital, which banks perceive as being costly.

In He et al. (2022), shocks to bond supply can be absorbed by two types of arbitrageurs: dealer banks, which are subject to non-risk-based equity capital requirements, and hedge funds, which are not. Hedge funds can finance their positions in bonds by borrowing from banks in the repo market. Banks' balance-sheet costs limit banks' willingness to hold positions in bonds and to lend to hedge funds in the repo market. In the presence of balance-sheet costs, shocks to bond supply push the spread between long and short rates and the spread between Treasury and overnight-index swap (OIS) yields—a failure of the Law of One Price that reflects bank balance-sheet costs—in the same direction. Balance-sheet costs also steepen the aggregate demand curve for short-rate risk, amplifying the impact of bond supply shocks on bond risk premia and yields. He et al. (2022) use the model to explain the sharp rise in long rates and the rise in Treasury yields relative to OIS yields during the COVID-19-induced "dash for cash" in March 2020.

A simple way to introduce balance-sheet costs in the model of Sections 2 and 3, and capture the key insights in He et al. (2022), is to assume that arbitrageurs are homogeneous and that balance-sheet costs enter directly into their objective function. Suppose that the arbitrageurs' objective function is

$$\mathbb{E}_t[\hat{W}_{t+1}] - \frac{a}{2}\mathbb{V}\mathrm{ar}_t[\hat{W}_{t+1}] - \frac{\psi}{2}(E_t)^2, \tag{98}$$

where E_t is equity capital and $\psi > 0$ is a constant that captures the notion that equity capital is costly for banks. Suppose additionally that arbitrageurs are subject to a binding equity capital requirement of the form

$$E_t = k \sum_{\tau=2}^{T} |X_t^{(\tau)}|, \tag{99}$$

where $k > 0$ is a constant. The constraint (99) requires that arbitrageurs hold equity capital against their bond positions for maturities 2 to T. We are excluding maturity 1 from the constraint (99), interpreting the one-period interest rate as a swap rate. Under current regulatory capital rules, interest-rate swaps indeed barely impact the amount of required equity capital. The absolute value in (99) captures the notion that banks are required to hold the same amount of equity capital against long and short positions in bonds. The constant k is approximately equal to 6% for U.S. banks.

For simplicity, we focus on the case of random and inelastic supply. We assume additionally that supply is always positive—i.e., $X_t^{(\tau)} = S_t^{(\tau)} > 0$ for all t and τ. This assumption ensures that arbitrageurs hold always long positions in all bonds in equilibrium and that the absolute-value constraint in (99) becomes linear.[10] We finally parameterize the function $\theta^{(\tau)}$ so an increase in the supply factor s_t raises aggregate bond supply for maturities 2 to T—i.e., $\sum_{\tau=2}^{T} \theta^{(\tau)} > 0$.

Arbitrageurs' first-order condition for τ-period bonds is

$$\mu_t^{(\tau)} - r_t = c_t + A_r^{(\tau-1)} \lambda_{r,t} + A_s^{(\tau-1)} \lambda_{s,t}, \tag{100}$$

where c_t is the marginal cost of equity capital, given by $c_t = \psi k^2 \sum_{\tau=2}^{T} (\zeta^{(\tau)} + \theta^{(\tau)} s_t)$, and $\lambda_{f,t}$ for $f = r, s$ are the prices of factor risk, given by (42). Equation (100) shows that in the presence of balance-sheet costs, expected excess bond returns reflect not only a compensation for risk but also a time-varying cost of equity capital c_t for arbitrageurs. An increase in c_t raises the expected returns on all bonds by the same amount irrespective of their maturity and, hence, of their risk.

Supply shocks have larger effects on bond yields in the presence of balance-sheet costs. This is because the shocks raise the marginal cost c_t of equity capital, thus raising expected bond returns and yields. Because the effect of supply shocks on bond yields is larger in the presence of balance-sheet costs, supply risk is also larger, and so is the persistence ρ_s^* of the supply factor under the risk-neutral measure. As a result, the effect of supply shocks in the presence of balance-sheet costs may be more pronounced for longer-term bonds. Countering this effect is the fact that, not being risk-based, shifts in balance-sheet costs can have large effects on short-term bonds. As a result, the effects of supply shocks on short-term yields can be more pronounced than those on long-term yields when supply shocks are transient.

The simple model presented above can also capture the effect of supply shocks on the

[10] Supply is always positive if the shocks $\varepsilon_{s,t}$ to the supply factor s_t are drawn from a distribution with bounded support and the function $\zeta^{(\tau)}$ is sufficiently positive.

spread between bond yields and swap rates. We treat swaps as if they require zero equity capital, which is a good approximation of current regulatory capital rules. While swaps do not entail balance-sheet costs, they still expose arbitrageurs to interest-rate risk. If arbitrageurs hold no swaps in equilibrium, then the equilibrium swap curve can be obtained by recomputing the equilibrium with $c_t = 0$. Thus, the model predicts that swap yields will be below government bond yields with the same maturity—i.e., swap spreads will be negative. Moreover, if dealers are required to absorb a large supply shock, as they arguably were during the March 2020 dash for cash, then the model predicts that both swap and Treasury yields will rise, but that Treasury yields will rise more.

Hanson et al. (2023) build a supply-and-demand driven model of swap spreads in the VV tradition where there are separate shocks to arbitrageurs' capital and to the bond supply that arbitrageurs must hold in equilibrium. The baseline model in that paper is affine, but the authors show how the theory can be extended to deal with the non-linearities that stem from balance-sheet constraints that bind only occasionally and from arbitrageurs' positions that switch sign over time.

6.6 Convenience yields

Starting with Krishnamurthy and Vissing-Jorgensen (2012), a large literature over the past decade has shown that U.S. government bonds and other safe securities may command a non-risk-based convenience yield. Convenience yields are straightforward to incorporate into the VV framework by adding non-pecuniary benefits of holding safe securities to the objective function of arbitrageurs. Suppose that the arbitrageurs' objective function is

$$\mathbb{E}_t[\hat{W}_{t+1}] - \frac{a}{2}\mathbb{V}\mathrm{ar}_t[\hat{W}_{t+1}] + m(X_t), \tag{101}$$

where $m(X_t)$ is a convenience benefit from holding an aggregate value $X_t \equiv \sum_{\tau=2}^{T} X_t^{(\tau)}$ of government bonds with maturities 2 to T. We interpret the one-period interest rate as a swap rate as in Section 6.5, and assume that swaps carry no convenience benefits. We assume that the function $m(X_t)$ is increasing and concave. For simplicity, we adopt the quadratic specification $m(X_t) = \alpha_m X_t - (\beta_m/2)X_t^2$.

As in Section 6.5, we focus on the case of random and inelastic supply. We assume that the aggregate supply $S_t \equiv \sum_{\tau=2}^{T} S_t^{(\tau)}$ for maturities 2 to T is always smaller than α_m/β_m, so that the marginal convenience benefit $m'(X_t) = m'(S_t) > 0$ is always positive in equilibrium. Arbitrageurs' first-order condition for τ-period bonds is

$$\mu_t^{(\tau)} - r_t = -m_t + A_r^{(\tau-1)}\lambda_{r,t} + A_s^{(\tau-1)}\lambda_{s,t}, \tag{102}$$

where m_t is the marginal convenience benefit, given by $m_t = \alpha_m - \beta_m \sum_{\tau=2}^{T}(\zeta^{(\tau)} + \theta^{(\tau)} s_t)$, and $\lambda_{f,t}$ for $f = r, s$ are the prices of factor risk, given by (42).

Convenience benefits and balance-sheet costs have opposite implications for the average level of yields but identical implications for the response of yields to supply shocks. Convenience benefits push bond yields down, while balance-sheet costs push yields up. However, both convenience benefits and balance-sheet costs lead supply shocks to have larger effects on bond yields than they otherwise would. And, both lead bond supply shocks to have larger effects on bond yields than on like-maturity swap rates. In the presence of balance sheet costs, this is because supply shocks raise the marginal cost c_t of equity capital. In the presence of convenience benefits, this is because supply shocks lower the marginal convenience benefit m_t. Formally, the function $A_s^{(\tau)}$ is the same in Sections 6.5 and 6.6 when $\psi k^2 = \beta_m$—i.e., when supply shocks raise c_t by the same amount as they lower m_t. However, convenience benefits and balance-sheet costs have opposite effects the function $C^{(\tau)}$.

6.7 Real versus nominal yields

Many developed countries issue both nominal and inflation-indexed bonds. Greenwood et al. (2023) show how the VV framework can be extended to model both nominal and real yield curves. Specifically, one would assume that there are separate, but potentially correlated, exogenous processes for the short-term real interest rate and for inflation. The process for the real rate is correlated with the process for inflation, following a standard Taylor rule. The model is closed by adding separate but potentially correlated shocks to the net supply of nominal and real bonds.

In such a model, the so-called break-even inflation rate—i.e., the difference between the yield on duration-matched nominal and real bonds—reflects both expected future inflation and the difference in risk premia between nominal and real bonds—i.e., an inflation risk premium. The inflation risk premium responds to differential net supply shocks between nominal and real bonds. For instance, holding expected inflation fixed, if the net supply of real bonds increases relative to the net supply of nominal bonds, this increases long-term real yields relative to nominal yields, pushing down break-even inflation. Such a model might prove useful in understanding the disruptions to the TIPS market that often seem to occur during periods of significant macro-financial distress such as the Fall of 2008 and March 2020 (see Campbell et al. (2009)). In both periods, real yields fell significantly less than nominal yields, so inflation break-even yields plummeted. Practitioner accounts suggest that these dynamics were driven by supply-demand imbalances, rather than expected deflation.

7 Other applications

In this section, we describe a number of further extensions that apply the VV framework beyond the domain of default-free bonds.

7.1 Foreign exchange

To study foreign exchange rates, Gourinchas et al. (2022) and Greenwood et al. (2023) both extend the VV framework to a two-country setting, say the U.S. and the Eurozone. There are short- and longer-term bonds in both countries as well as an exchange rate between the two countries' currencies. The short rate in each country follows an exogenous AR(1) processes and the two short rates can be correlated. As in VV, specialized bond arbitrageurs must absorb random shocks to the supply of long-term bonds in each country as well as random shocks to the supply of foreign exchange. Gourinchas et al. (2022) work in continuous time and consider a continuum of long-term bonds in each country. Greenwood et al. (2023) work in discrete time and consider a single class of perpetual long-term bonds in each country. However, the underlying economics is similar.

These models predict that shifts in the supply of long-term bonds impact not only bond term premia, but also the expected returns on the foreign exchange (FX) trade that borrows in dollars and lends in euro. For example, an increase in the supply of long-term U.S. bonds raises both the expected excess return on long-term U.S. bonds and the expected return on the FX trade, leading to a depreciation of the euro versus the dollar. Conversely, reductions in the supply of long-term U.S. bonds, such as those that result from quantitative easing, will lead the dollar to depreciate against the euro, matching recent evidence in Bauer and Neely (2014), Neely (2015), Swanson (2017), and Bhattarai and Neely (2022).

The intuition for this result is that long-term U.S. bonds and the borrow-in-dollar lend-in-euro FX trade have similar exposures to U.S. short-rate risk. Specifically, when the U.S. short rate rises unexpectedly, (i) EH logic implies that yields on long-term U.S. bonds rise and their prices fall and (ii) uncovered-interest-rate-parity logic implies that the euro depreciates against the dollar so the borrow-in-dollar lend-in-euro FX trade also suffers losses.

Consider now the effect of an increase in the net supply of long-term U.S. bonds. Following this supply shift, bond arbitrageurs become more exposed to future shocks to the U.S. short rate. As a result, the price of bearing U.S. short-rate risk rises. Since long-term U.S. bonds are exposed to U.S. short-rate risk, this leads to a rise in the risk-premium component of long-term U.S. yields. It also leads to a rise in the risk premium on the borrow-in-dollar lend-in-euro FX trade, which is similarly exposed to U.S. short-rate risk. As a result, the euro depreciates against the dollar and is expected to appreciate going forward.

These models make several additional predictions. First, bond supply shocks should have a larger impact on the bilateral exchange rate when the correlation between the two countries' short rates is lower. For example, the JPY-USD exchange rate should be more responsive to U.S. QE than the EUR-USD exchange rate because Japanese short rates are less correlated with U.S. short rates than are Euro short rates. Second, these models match the otherwise puzzling finding in Lustig et al. (2019) that the return to the FX trade declines if one borrows long-term in one currency to lends long-term in the other. This pattern arises because the "long-term" FX trade has offsetting exposures to short-rate shocks, making it less risky for arbitrageurs than the standard FX trade involving short-term bonds. Third, if one assumes that the supply of bonds and the supply of foreign exchange are increasing in their price, these models offer a unified explanation that links the predictability of FX returns documented by Fama (1984)—i.e., a larger short-rate differential predicts higher FX returns—with the predictability of long-term bond returns documented by Fama and Bliss (1987) and Campbell and Shiller (1991)—i.e., a steeper yield curve predicts higher bond excess returns.

7.2 Credit risk

There are several ways to introduce credit risk into VV-style models. Greenwood et al. (2018) develop an approach that is appropriate for modelling a *diversified portfolio* of defaultable bonds where it is reasonable to assume that portfolio-level default losses follow an AR(1) process in discrete time or an OU process in continuous time. However, their approach is not appropriate for modelling the term structure of yields for an individual corporate or sovereign borrower where default is a binary event. Costain et al. (2022) outline an approach that is appropriate for individual borrowers where a binary default event arrives at some hazard rate. Focusing on the Eurozone, they model the term structures of two sovereigns that are subject to the same movements in riskless short rates—i.e., from the European Central Bank. However, the term structure for one sovereign is free of default risk, while the term structure for the other is subject to the arrival of Poisson default events.[11] In both frameworks, both default-free and defaultable bonds are exposed to short-rate shocks, but only defaultable bonds are exposed to default losses. Thus, shocks to the supply of either default-free or defaultable bonds will shift the market price of short-rate risk. However, only shocks to the supply of defaultable bonds will shift the market price of default risk. Gilchrist et al. (2021) study the impact of the Federal Reserve's purchases of corporate bonds.

7.3 Further market segmentation

There is segmentation in the VV model in the sense that the marginal investors in bonds are specialized bond arbitrageurs. However, further segmentation is possible—particularly segmentation across different asset classes within fixed-income markets—which has implications for how supply shocks impact prices. Segmentation of this sort is explored in Greenwood et al. (2018) and Greenwood et al. (2023).

Suppose there are two classes of long-term bonds—say, government bonds that are free of default risk and corporate bonds that are exposed to default risk. Suppose that not

[11] The default hazard rate can evolve deterministically over time. It cannot evolve stochastically if the model is to remain affine.

all bond arbitrageurs are able and willing to substitute between all assets. Specifically, a fraction π of bond arbitrageurs are highly specialized. Of these specialists, a fraction ϕ specialize in government bonds and only substitute between short-term and longer-term government bonds, and a fraction $1-\phi$ specialize in corporate bonds and only substitute between short-term government bonds and longer-term corporate bonds. Fraction $1-\pi$ of bond arbitrageurs are generalists who substitute between short-term government bonds and both types of long-term bonds.

When $\pi = 1$ the government and corporate bond markets are completely segmented, when $\pi = 0$ they are fully integrated, and when $\pi \in (0, 1)$ they are partially segmented. In this setting, one can show that own-market price impact—i.e., the impact of a shock to the supply of an asset class on prices of that asset class—is greatest when markets are highly segmented (approaching $\pi = 1$) and declines as markets become more highly integrated (approaching $\pi = 0$). Cross-market spillovers of supply shocks—e.g., the way that corporate bonds react to shocks to the supply of government bonds—depend on two key factors. First, on the degree of fundamental substitutability between the two asset classes as captured by their covariance. Second, on the degree of segmentation between the two markets. Specifically, spillovers are greatest when the markets are tightly integrated and all arbitrageurs are able and willing to substitute between the two asset classes (approaching $\pi = 0$). Spillovers decline if some arbitrageurs do not substitute between the two asset classes (increasing π).

8 Conclusion

In this review, we develop a model of the term structure based on supply and demand. After laying out the main results and intuitions, we show that the model can help explain a number of empirical findings. In addition, the model can be adapted to handle a number of extensions, including forward guidance, wealth effects, convenience yields, exchange rates and credit risk. Our hope is that this presentation can help researchers adapt the VV model, or ones similar to it, in other settings.

While the VV approach has been widely adopted in finance settings, it has made more limited inroads into macroeconomics. Consider the question of how QE affects the real economy, as noted by Woodford (2016). The VV approach can rationalize why a reduction in net bond supply held by the public leads to a flattening of the yield curve. But it is another step to go from the impact on long-term bond yields to an account of how QE impacts firm investment, household consumption, and the broader real economy. Similarly, consider the case of QE by the European Central Bank leading to depreciation of the Euro relative to the dollar, as noted by Gourinchas et al. (2022) and Greenwood et al. (2023). What are the corresponding implications for exports and total output, and what are the feedback loops into bond prices? Ray (2019) and Ray et al. (2023) make progress on some of these questions by embedding a VV segmented bond markets block into a New Keynesian model. Macroeconomic models with bond market segmentation also include Andres et al. (2004) and Sims et al. (2023). Much more remains to be done, both in characterizing how financial shocks affect the macroeconomy when capital markets are segmented, and in deriving optimal monetary and fiscal policy in these settings.

References

Andres, J., Lopez-Salido, D., and Nelson, E. (2004). Tobin's imperfect asset substitution in optimizing general equilibrium. *Journal of Money, Credit and Banking*, 36(4):665–690.

Bauer, M. D. and Neely, C. J. (2014). International channels of the Fed's unconventional monetary policy. *Journal of International Money and Finance*, 44:24–46.

Bhattarai, S. and Neely, C. J. (2022). An Analysis of the Literature on International Unconventional Monetary Policy. *Journal of Economic Literature*, 60(2):527–597.

Campbell, J. Y. and Shiller, R. J. (1988). Stock prices, earnings, and expected dividends. *The Journal of Finance*, 43(3):661–676.

Campbell, J. Y. and Shiller, R. J. (1991). Yield Spreads and Interest Rate Movements: A Bird's Eye View. *The Review of Economic Studies*, 58(3):495–514. Publisher: [Oxford University Press, Review of Economic Studies, Ltd.].

Campbell, J. Y., Shiller, R. J., and Viceira, L. M. (2009). Understanding Inflation-Indexed Bond Markets.

Carboni, G. and Ellison, M. (2022). Preferred habitat and monetary policy through the looking-glass. Working Paper Series 2697, European Central Bank.

Cochrane, J. (2008). Comments on 'Bond Supply and Excess Bond Returns' by Robin Greenwood and Dimitri Vayanos. Working paper, University of Chicago.

Costain, J., Nuño, G., and Thomas, C. (2022). The Term Structure of Interest Rates in a Heterogeneous Monetary Union. *CESifo Working Paper Series*. Number: 9844 Publisher: CESifo.

Culbertson, J. (1957). The term structure of interest rates. *Quarterly Journal of Economics*, 71:485–517.

Domanski, D., Shin, H. S., and Sushko, V. (2017). The Hunt for Duration: Not Waving but Drowning? *IMF Economic Review*, 65(1):113–153. Publisher: [Palgrave Macmillan Journals, International Monetary Fund].

Duffie, D. (2010). Presidential Address: Asset Price Dynamics with Slow-Moving Capital. *The Journal of Finance*, 65(4):1237–1267. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1540-6261.2010.01569.x.

D'Amico, S. and King, T. B. (2013). Flow and stock effects of large-scale treasury purchases: Evidence on the importance of local supply. *Journal of Financial Economics*, 108(2):425–448.

Eggertsson, G. B. and Woodford, M. (2003). The Zero Bound on Interest Rates and Optimal Monetary Policy. *Brookings Papers on Economic Activity*, 34(1):139–235.

Fama, E. F. (1984). Forward and spot exchange rates. *Journal of Monetary Economics*, 14(3):319–338.

Fama, E. F. and Bliss, R. R. (1987). The Information in Long-Maturity Forward Rates. *The American Economic Review*, 77(4):680–692. Publisher: American Economic Association.

Fratzscher, M., Lo Duca, M., and Straub, R. (2018). On the International Spillovers of US Quantitative Easing. *The Economic Journal*, 128(608):330–377. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ecoj.12435.

Gagnon, J., Raskin, M., Remache, J., and Sack, B. (2011). The Financial Market Effects of the Federal Reserve's Large-Scale Asset Purchases. *International Journal of Central Banking*, 7(1):3–43. Publisher: International Journal of Central Banking.

Garbade, K. D. and Rutherford, M. (2007). Buybacks in Treasury cash and debt management. Staff Reports 304, Federal Reserve Bank of New York.

Gilchrist, S., Wei, B., Z Yue, V., and Zakrajšek, E. (2021). The fed takes on corporate credit risk: an analysis of the efficacy of the smccf. BIS Working Papers 963, Bank for International Settlements.

Gourinchas, P.-O., Ray, W. D., and Vayanos, D. (2022). A Preferred-Habitat Model of Term Premia, Exchange Rates, and Monetary Policy Spillovers.

Greenwood, R., Hanson, S., Stein, J. C., and Sunderam, A. (2023). A Quantity-Driven Theory of Term Premia and Exchange Rates*. *The Quarterly Journal of Economics*, 138(4):2327–2389.

Greenwood, R., Hanson, S. G., and Liao, G. Y. (2018). Asset Price Dynamics in Partially Segmented Markets. *The Review of Financial Studies*, 31(9):3307–3343. Publisher: [Oxford University Press, The Society for Financial Studies].

Greenwood, R., Hanson, S. G., and Vayanos, D. (2016). Forward Guidance in the Yield Curve: Short Rates versus Bond Supply. *Central Banking, Analysis, and Economic Policies Book Series*, 24:11–62. Publisher: Central Bank of Chile.

Greenwood, R. and Vayanos, D. (2010). Price Pressure in the Government Bond Market. *American Economic Review*, 100(2):585–590.

Greenwood, R. and Vayanos, D. (2014). Bond Supply and Excess Bond Returns. *The Review of Financial Studies*, 27(3):663–713.

Greenwood, R. M. and Vissing-Jorgensen, A. (2018). The Impact of Pensions and Insurance on Global Yield Curves.

Haddad, V. and Sraer, D. (2020). The Banking View of Bond Risk Premia. *The Journal of Finance*, 75(5):2465–2502. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jofi.12949.

Hanson, S. G. (2014). Mortgage convexity. *Journal of Financial Economics*, 113(2):270–299.

Hanson, S. G., Lucca, D. O., and Wright, J. H. (2021). Rate-Amplifying Demand and the Excess Sensitivity of Long-Term Rates*. *The Quarterly Journal of Economics*, 136(3):1719–1781.

Hanson, S. G., Malkhozov, A., and Venter, G. (2023). Demand-and-Supply Imbalance Risk and Long-Term Swap Spreads.

Hanson, S. G. and Stein, J. C. (2015). Monetary policy and long-term real rates. *Journal of Financial Economics*, 115(3):429–448.

Hayashi, F. (2018). Computing equilibrium bond prices in the vayanos-vila model. *Research in Economics*, 72(2):181–195.

He, Z., Nagel, S., and Song, Z. (2022). Treasury inconvenience yields during the COVID-19 crisis. *Journal of Financial Economics*, 143(1):57–79.

Joyce, M. A. S., Lasaosa, A., Stevens, I., and Tong, M. (2011). The Financial Market Impact of Quantitative Easing in the United Kingdom. *International Journal of Central Banking*, 7(3):113–161. Publisher: International Journal of Central Banking.

Kekre, R., Lenel, M., and Mainardi, F. (2023). Monetary Policy Segmentation, and the Term Structure.

King, T. B. (2019). Expectation and duration at the effective lower bound. *Journal of Financial Economics*, 134(3):736–760.

Krishnamurthy, A. and Vissing-Jorgensen, A. (2012). The Aggregate Demand for Treasury Debt. *Journal of Political Economy*, 120(2):233–267. Publisher: University of Chicago Press.

Lustig, H., Stathopoulos, A., and Verdelhan, A. (2019). The Term Structure of Currency Carry Trade Risk Premia. *American Economic Review*, 109(12):4142–4177.

Malkhozov, A., Mueller, P., Vedolin, A., and Venter, G. (2016). Mortgage Risk and the Yield Curv. *The Review of Financial Studies*.

Modigliani, F. and Sutch, R. (1966). Innovations in Interest Rate Policy. *The American Economic Review*, 56(1/2):178–197. Publisher: American Economic Association.

Neely, C. J. (2015). Unconventional monetary policy had large international effects. *Journal of Banking & Finance*, 52:101–111.

Ray, W. (2019). Monetary Policy and the Limits to Arbitrage: Insights from a New Keynesian Preferred Habitat Model. *2019 Meeting Papers*. Number: 692 Publisher: Society for Economic Dynamics.

Ray, W., Droste, M., and Gorodnichenko, Y. (2023). Unbundling Quantitative Easing: Taking a Cue from Treasury Actions. *Journal of Political Economy (Forthcoming)*.

Sims, E., Wu, J. C., and Zhang, J. (2023). The four-equation new keynesian model. *The Review of Economics and Statistics*, 105(4):931–947.

Tobin, J. (1958). Liquidity Preference as Behavior Towards Risk. *The Review of Economic Studies*, 25(2):65–86. Publisher: [Oxford University Press, Review of Economic Studies, Ltd.].

Tobin, J. (1969). A General Equilibrium Approach To Monetary Theory. *Journal of Money, Credit and Banking*, 1(1):15–29. Publisher: [Wiley, Ohio State University Press].

Vayanos, D. and Vila, J.-L. (2009). A Preferred-Habitat Model of the Term Structure of Interest Rates.

Vayanos, D. and Vila, J.-L. (2021). A Preferred-Habitat Model of the Term Structure of Interest Rates. *Econometrica*, 89(1):77–112. Publisher: Econometric Society.

Wiliams, J. (2014). Monetary policy at the zero lower bound: Putting theory into practice. In *Hutchins Center on Fiscal and Monetary Policy*. Brookings Institution.

Woodford, M. (2016). Quantitative easing and financial stability. 24:151–233.