Received: 14 March 2019 | Revised: 4 December 2020 | Accepted: 4 December 2020

DOI: 10.1002/jae.2828

RESEARCH ARTICLE

WILEY Journal of APPLIED ECONOMETRICS

No-arbitrage priors, drifting volatilities, and the term structure of interest rates

Andrea Carriero[1,2] | **Todd E. Clark**[3] | **Massimiliano Marcellino**[4]

[1]Department of Economics, Queen Mary, University of London, London, UK

[2]Department of Economics, University of Bologna, Bologna, Italy

[3]Research Department, Federal Reserve Bank of Cleveland, Cleveland, Ohio, USA

[4]Department of Economics, Bocconi University, IGIER and CEPR, Milan, Italy

Correspondence
Massimiliano Marcellino, Department of Economics, Bocconi University, IGIER and CEPR, Milan, Italy.
Email: massimiliano.marcellino@unibocconi.it

Abstract

We use a Bayesian vector autoregression with stochastic volatility to forecast government bond yields. We form the conjugate prior from a no-arbitrage affine term structure model. The model improves on the accuracy of point and density forecasts from a no-change random walk and an affine term structure model with stochastic volatility. Our proposed approach may succeed by relaxing the no-arbitrage affine term structure model's requirements that yields obey a factor structure and that the factors follow a Markov process. In the term structure model, its cross-equation no-arbitrage restrictions on the factor loadings appear to play a marginal role in forecasting gains.

KEYWORDS

density forecasting, no arbitrage, term structure, volatility

1 | INTRODUCTION

Producing accurate forecasts of the term structure of interest rates is crucial for bond portfolio management, derivatives pricing, and risk management. In the recent literature, several papers have analyzed the forecasting performance of different methods, for example, Almeida and Vicente (2008), Ang and Piazzesi (2003), Carriero (2011), Carriero et al. (2012), Christensen et al. (2011), Diebold and Li (2006), and Duffee (2002). All of these contributions have focused on point forecasts of the yield curve, but assessing the whole predictive density of the yield curve is more important for the success of portfolio and risk management strategies. Hong et al. (2004) make a relevant contribution in this context, finding that modeling changes in the variance is important for interest rate density forecasting. Shin and Zhong (2017) find a similar result using realized volatility.

In this paper, we focus on both point and density forecasting of government bond yields, using a vector autoregression (VAR henceforth) with two key features. First, based on the observation that both yields and their volatilities strongly co-move over time, we assume that the conditional volatilities of the yields are time varying and driven by a latent common volatility factor. Second, conditionally on the volatilities, the VAR is Gaussian, which permits specifying a conjugate prior on the VAR coefficients. We choose to parameterize the conjugate prior in a way that reflects the term structure model, by centering it around a specific theoretical model for the yield curve. Specifically, we use the canonical affine term structure model (ATSM) of Duffie and Kan (1996), in its equivalent but computationally more stable representation developed by Joslin et al. (2011).

The rationale for using an ATSM as a prior on a more general model lies in the somewhat disappointing performance that ATSMs have shown in forecasting yields out of sample, documented in contributions such as Duffee (2002) and Ang and Piazzesi (2003), which both show that these models cannot beat a random walk forecast. One reason behind this finding might be that beyond the mere assumption of the absence of arbitrage—which is per se reasonable in well-developed

markets—these models require a set of additional specification assumptions, which do not necessarily hold in the data and therefore introduce misspecification (see, e.g., Hamilton & Wu (2014)). For example, yields are assumed to follow a factor model, with the factor loadings obeying a set of complex nonlinear restrictions and with the factors following a Markov process (i.e., they depend exclusively on their own value in the previous period).

Using the no-arbitrage model restrictions to form a prior for a VAR rather than imposing the restrictions sharply allows us to take into account its potential misspecification. The prior will shrink the estimated coefficients of the VAR—and in turn the density forecasts—in an economically meaningful direction, which might (and in our application does) improve the forecasting performance with respect to both a no-arbitrage model and an unrestricted VAR. This can be achieved in a homoskedastic setting by using the hierarchical prior approach proposed in Del Negro and Schorfheide (2004).[1,2]

This paper extends the methodology of Del Negro and Schorfheide (2004) to the case of VARs featuring stochastic volatility. As the volatilities of a panel of yields move closely together, we impose a factor structure where the volatility of each yield is related to a common stochastic volatility (CSV) factor, as in Carriero et al. (2016). Hence, our approach shrinks point and density forecasts toward values consistent with a no-arbitrage term structure model, while also allowing for time variation in the volatilities. Our approach is also an extension of that of Giannone et al. (2015) to the case in which a hierarchical specification is used in not only the prior mean but also the prior variance. Our results show that a hierarchical specification for the prior means does help in forecasting, suggesting that not only shrinkage per se but also the direction of shrinkage can be helpful.

Estimation of the VAR model using US data on government bond yields covering the period from January 1985 to December 2018 shows that the proposed approach produces better density forecasts than an ATSM with time-varying volatility along the lines of Hautsch and Ou (2012). Compared with a random walk, which is typically a very strong benchmark in forecasting the yields, the proposed model fares consistently better at the short and medium end of the curve and equally well at the long end of the curve.

Further analysis reveals that the approach we propose might work better than a term structure model because it relaxes the requirements that yields obey a strict factor structure and that the factors follow a Markov process. Instead, we find that the cross-equation no-arbitrage restrictions on the factor loadings only play a marginal role, in line with Duffee (2011a).[3]

A number of papers have carried out Bayesian estimation of dynamic term structure models. Examples include Ang et al. (2011), Bauer (2018), Chib and Ergashev (2009), Chib and Kang (2014), Creal and Wu (2015, 2017), and Hautsch and Ou (2012). This paper contributes to this literature. Still, the method proposed can be applied to a wide range of alternative models, the only requirement being that they admit a Gaussian linear state-space representation.

The paper is organized as follows. Section 2 describes the no-arbitrage model used as a prior. Section 3 discusses the proposed approach, derives the conditional posteriors, and describes an Markov chain Monte Carlo (MCMC) sampler for estimation (with details in Appendix S1). Section 4 presents our US-based evidence, including both a full-sample evaluation and an out-of-sample forecasting assessment. Section 5 summarizes the main results and concludes. Additional details and results can be found in Appendix S1.

2 | THE ATSM

Since the seminal work of Vasicek (1977), a large part of research has focused on Gaussian ATSMs. Prominent contributions in this tradition include Ang and Piazzesi (2003), Dai and Singleton (2000), Duffee (2002), and Duffie and Kan (1996). Traditional ATSMs entail a high level of nonlinearity that makes the estimation extremely difficult and often unreliable, an issue discussed extensively in Duffee and Stanton (2012) and Hamilton and Wu (2012). Some recent literature has successfully addressed this problem. Hamilton and Wu (2012) proposed a strategy to estimate such models using a series of

[1]In the Del Negro and Schorfheide (2004) approach, all model coefficients and latent variables—both of the VAR and of the economic model used as a prior—are estimated jointly, and their posterior distributions are shrunk in the economically meaningful direction. The approach is, in spirit, similar to the relative entropy procedures of Robertson et al. (2005) and Giacomini and Ragusa (2014). In the entropy approach, the forecasts are "tilted" toward an economic model of reference after estimation of a baseline (atheoretical) model has taken place, and the parameters of the economic model of reference need to be estimated separately.

[2]Carriero (2011) conducts this exercise in a homoskedastic setting and shows that once the misspecification of the model is properly taken into account, the point forecasting performance can improve substantially. However, he works under the hypotheses of homoskedasticity of the yields, a mild assumption for point forecasting but likely inadequate for density forecasting. Moreover, whereas Carriero's (2011) prior specification is based on the model by Ang and Piazzesi (2003), here, we consider the new canonical form of no-arbitrage models introduced by Joslin et al. (2011), which presents important advantages in the computation of the likelihood, providing the MCMC sampler with better mixing properties and faster computation time.

[3]Duffee (2011a) shows that the loadings of a Gaussian ATSM can be estimated with extremely high precision without imposing the no-arbitrage restrictions.

transformations and ordinary least squares (OLS) estimation. Christensen et al. (2011) proposed a model based on the Nelson and Siegel (1987) exponential framework, which still imposes no arbitrage but can be estimated more reliably.

In this paper, we use the representation proposed by Joslin et al. (2011) (JSZ henceforth), which is equivalent to the canonical representation of Duffie and Kan (1996) but parameterized in such a way that estimation is considerably simplified. Let y_t denote an N-dimensional vector of yields on a set of zero-coupon bonds of maturity $\tau = 1, \dots, N$. In the JSZ representation, yields are driven by an n-dimensional vector of unobservable risk factors P_t:

$$y_t = A_P + B_P P_t + \Sigma_y \varepsilon_t^y, \tag{1}$$

$$P_t = K_{0P}^{\mathbb{P}} + (K_{1P}^{\mathbb{P}} + I_n) P_{t-1} + \Sigma_P \varepsilon_t^{\mathbb{P}}, \tag{2}$$

where A_P and B_P are $N \times 1$ and $N \times n$ coefficient matrices, $K_{0P}^{\mathbb{P}}$ is an $n \times 1$ vector, $K_{1P}^{\mathbb{P}}$ is an $n \times n$ matrix, and Σ_y and Σ_P are lower triangular Cholesky factor matrices. The disturbance vectors ε_t^y and $\varepsilon_t^{\mathbb{P}}$ contain, respectively, N and n univariate, mutually independent, *i.i.d.*$N(0, 1)$ stochastic processes. The subscript P indicates that an object belongs to the JSZ model, which is based on the factors P_t. The superscript $\mathbb{P}$ indicates an object specified under the physical (i.e., real world) measure of probability.

The advantage of the JSZ representation stems from the fact that the unobservable factors P_t can be easily approximated by an observable portfolio of yields $P_t^o = W y_t$ where W contains the loadings of the first n principal components of y_t. The approximation is such that a least squares regression of P_t^o on P_{t-1}^o will recover the maximum likelihood estimates of $K_{0P}^{\mathbb{P}}$ and $K_{1P}^{\mathbb{P}}$, which means that these parameters can be estimated in a preliminary step and concentrated out of the likelihood.[4] Moreover, the observable P_t^o provides a straightforward initial condition for the filtering of the unobservable states P_t and the estimation of Σ_P. Details on how the representation (1 and 2) is equivalent to the representation of Duffie and Kan (1996) can be found in Appendix S1, Section A.

Equations (1) and (2) constitute a factor model in which the yields depend linearly on the factors P_t through the intercept vector A_P and the factor loadings B_P. The assumption of the absence of arbitrage entails an internal consistency across yields of different maturities, which in turn implies that the elements A_P and B_P must obey a set of (highly) nonlinear restrictions. Let $A_P = (I - B_P W) A_S$ and $B_P = B_S (W B_S)^{-1}$, where A_S and B_S are an $n \times 1$ vector and an $n \times n$ matrix, and let $A_S(\tau)$ and $B_S(\tau)$ denote the elements in the τth row of these objects (recall that τ denotes the maturity). The no-arbitrage restrictions are

$$A_S(\tau) = -A_\tau / \tau,\ A_{\tau+1} = A_\tau + K_{0S}^{\mathbb{Q}\prime} B_\tau + 0.5 B_\tau' \Sigma_S \Sigma_S' B_\tau, \tag{3}$$

$$B_S(\tau) = -B_\tau / \tau,\ B_{\tau+1} = B_\tau + K_{1S}^{\mathbb{Q}\prime} B_\tau - i, \tag{4}$$

with initial condition $A_0 = 0$ and $B_0 = 0$. In the expressions above, $K_{0S}^{\mathbb{Q}} = (k_\infty^{\mathbb{Q}}, 0, \dots, 0)$ is an $n \times 1$ vector, $K_{1S}^{\mathbb{Q}} = J(\lambda^{\mathbb{Q}})$ is an $n \times n$ matrix in Jordan form with eigenvalues $\lambda^{\mathbb{Q}}$, Σ_S is a lower triangular Cholesky factor, and i is an $n \times 1$ vector of ones.

The subscript S indicates that an object belongs to the equivalent Duffie and Kan (1996) canonical representation of the model, which is the one in which (3) and (4) are expressed and which is based on an alternative set of factors S_t. The factors S_t are a rotation of the factors P_t, but they do not have a straightforward observable counterpart.[5]

The Riccati equations (3 and 4) are expressed as a function of the matrices $K_{0S}^{\mathbb{Q}}$ and $K_{1S}^{\mathbb{Q}}$, which are matrices describing the evolution of the state variables under the so-called equivalent martingale measure $\mathbb{Q}$. This is a probability measure corresponding to a hypothetical situation in which investors are risk neutral, as opposed to being risk averse as happens in the physical world described by the measure $\mathbb{P}$. Recall that the matrices $K_{0P}^{\mathbb{P}}$ and $K_{1P}^{\mathbb{P}}$ appearing in (2) describe the evolution of the state variables under the physical measure $\mathbb{P}$. Under the physical measure, agents' risk aversion implies that prices need to be predictable to some extent, producing the expected returns necessary to compensate investors for bearing risks.

[4] That is, we estimate these parameters in a preliminary step using a OLS regression and observable factors estimated with the principal components P_t^o. Strictly speaking, this concentration is exact only if one assumes that the P_t is observable without error. However, as noted by JSZ, the choice of principal components weights ensures that $P_t^o \approx P_t$, and the concentration is nearly exact.

[5] The relation between the factors of the two representations is $P_t = W A_P + W B_P S_t$.

The existence of the equivalent martingale measure $\mathbb{Q}$ is a necessary and sufficient condition for the absence of arbitrage. Conversion from the $\mathbb{P}$ to the $\mathbb{Q}$ measure can be achieved using a variable transformation described by a Radon–Nikodym derivative.[6]

It is important to distinguish the assumption of the absence of arbitrage and the additional specification restrictions inherent in an ATSM. For example, additional restrictions include the choice of modeling the yields with a factor structure as in (1) and the assumption that the factors follow the process in (2), which only features one lag and therefore is a Markov process. There is no guarantee that these additional assumptions are satisfied, and contributions such as Duffee (2011b), Joslin et al. (2014), and Joslin et al. (2013a, 2013b) have shown evidence to the contrary.[7]

The model at hand is a linear Gaussian state-space system with unobservable states P_t. As we have mentioned above, the JSZ representation makes it possible to concentrate out of the likelihood the coefficient matrices $K_{0P}^{\mathbb{P}}$ and $K_{1P}^{\mathbb{P}}$ appearing in (2) by estimating them via OLS. Moreover, the coefficient matrices $K_{0S}^{\mathbb{Q}} = (k_\infty^{\mathbb{Q}}, 0, \dots, 0)$ and $K_{1S}^{\mathbb{Q}} = J(\lambda^{\mathbb{Q}})$ appearing in (3) and (4) are parameterized by the scalar $k_\infty^{\mathbb{Q}}$ and the eigenvalue vector $\lambda^{\mathbb{Q}}$.[8] Therefore, the model is parameterized by

$$\theta = (\lambda^{\mathbb{Q}}, k_\infty^{\mathbb{Q}}, \Sigma_P, \Sigma_y). \tag{5}$$

For future reference, it is convenient to compute the moments of the yields for a given parameterization θ. Conditional on a given θ, the moments of y_t implied by the state-space system (1 and 2) are

$$E[y_t y_t'] = (A_p + B_p\bar{P})(A_p + B_p\bar{P})' + B_p\Sigma_f B_p' + \Sigma_y\Sigma_y', \tag{6}$$

$$E[y_t y_{t-h}'] = (A_p + B_p\bar{P})(A_p + B_p\bar{P})' + B_p(K_{1P}^P + I)^h\Sigma_f B_p', \tag{7}$$

where $\bar{P} = E[P_t] = -K_{1P}^{\mathbb{P}\,-1}K_{0P}^{\mathbb{P}}$ and Σ_f solves the Lyapunov equation $\Sigma_f = (K_{1P}^{\mathbb{P}} + I)\Sigma_f(K_{1P}^{\mathbb{P}} + I)' + \Sigma_P\Sigma_P'$. See Appendix S1, Section B for a derivation. Importantly, the system (6 and 7) is linear and Gaussian, with mean and variance being sufficient statistics, and this is what permits using it to form a conjugate prior.

The next section will illustrate how the moments (6) and (7) can be used to form a prior for a VAR. Because the JSZ model is Gaussian and homoskedastic, the prior will also be Gaussian and homoskedastic. Alternatively, one could think of using a prior that is already based on a model featuring drifting volatilities. To that end, there are broad classes of no-arbitrage models with stochastic volatility available, which could be potentially used as a prior. However, without Gaussianity, the moments (6) and (7) would no longer be sufficient statistics. While using a heteroskedastic model as the prior is in principle possible in our setup, the implementation would require repeated simulation of artificial datasets, which is in practice unmanageable.

3 | THE FULL HIERARCHICAL MODEL

As discussed in the previous section, any ATSM needs additional specification assumptions—beyond the mere absence of arbitrage—in order to be estimated and made operational. Therefore, it will suffer from misspecification to some degree. Enough misspecification may overwhelm any gains from parsimony and harm forecast accuracy. Instead, a VAR—provided its dynamics are sufficiently rich—is more likely to offer an accurate representation of the data. We model the yields using a VAR, while at the same time shrinking the VAR parameters in the direction implied by the JSZ model. This pushes the estimates toward the reference model, but it is less likely to create enough misspecification to harm forecast accuracy. In this section, we first describe the baseline general VAR model with CSV (VAR-CSV). Then, we discuss the likelihood, priors, and conditional posteriors.

[6]In particular, under the $\mathbb{Q}$ measure, the price of an asset V_t that does not pay any dividends at time $t+1$ satisfies $V_t = E_t^{\mathbb{Q}}[\exp(-r_t)V_{t+1}]$, where r_t is the short-term rate. Under the $\mathbb{P}$ measure, the price is $V_t = E_t^{\mathbb{P}}[(\xi_{t+1}/\xi_t)\exp(-r_t)V_{t+1}]$, where ξ_{t+1} is the Radon–Nikodym derivative.

[7]Duffee (2011b) shows that it is entirely possible for the factors to follow richer dynamics and that this translates to the presence of hidden factors that—although not useful in explaining the cross-section of yields—can help in explaining their dynamics. Similarly, Joslin et al. (2014) and Joslin et al. ((2013a), (2013b)) show that a more general model for the factors, including measures of real economic activity and inflation or more lags, better captures the dynamics of the term structure.

[8]The subscript ∞ emphasizes the long-run interpretation of this parameter. When the model is stationary under the $\mathbb{Q}$ measure, $k_\infty^{\mathbb{Q}}$ is proportional to the risk-neutral long-run mean of the short rate, and the model can be equivalently parameterized in terms of either parameter.

3.1 | A VAR model with CSV

Consider the autoregression of the $N \times 1$ vector of observable yields y_t:

$$y_t = \Phi_0 + \Phi_1 y_{t-1} + \ \dots \ + \Phi_p y_{t-p} + u_t, \tag{8}$$

where Φ_0 is an $N \times 1$ vector of intercepts and $\Phi_1, \ \dots \ , \Phi_p$ are $N \times N$ matrices of lagged coefficients. The $N \times 1$ vector of disturbances u_t is a mixture of normals:

$$u_t = \lambda_t^{0.5} \epsilon_t, \ \epsilon_t \sim N(0, V), \tag{9}$$

where $\lambda_t^{0.5}$ is a scalar latent variable evolving according to the log-normal process

$$\log(\lambda_t) = \phi_0 + \phi_1 \log(\lambda_{t-1}) + \nu_t, \ \nu_t \sim \text{i.i.d.} N(0, \phi_2). \tag{10}$$

We group the parameters governing the dynamics of λ_t in the vector $\phi = [\phi_0, \phi_1, \phi_2]$, and we set the initial condition $\lambda_1 = 1$ in order to achieve identification of the error variance matrix $Var(u_t) \equiv \Sigma_t = \lambda_t V$.

The model described in (8)–(10) is a VAR-CSV. Each variable in the model is a yield with a different maturity, but there is a single stochastic volatility process λ_t that is common to all yields, and drives the time variation in the entire variance-covariance matrix of the disturbances. This specification was proposed by Carriero et al. (2016) for macroeconomic variables.

The assumption of CSV is predicated on the fact that the volatilities of yields feature a strong factor structure, with the first principal component explaining most of the variation in the panel. For example, in the dataset we use in our empirical application, there is a strong commonality, with the first principal component explaining 89% of the individual volatilities of the yields.[9] As we shall see, the assumption of a single factor makes it possible to use a conjugate prior and to adapt the approach proposed by Del Negro and Schorfheide (2004). Of course, such an assumption does not come without a cost, and there are some reasons to believe that the assumption of a single factor might be too restrictive.[10] However, even with one factor only, the proposed model is more general and improves over homoskedastic specifications, such as the one of Carriero (2011).

3.2 | Likelihood

Consider the equations for all observations $t = 1, \ \dots \ , T$ and let Y be the $T \times N$ matrix with rows y_t', let X be the $T \times k$ matrix with rows $x_t' = [1, y_{t-1}', y_{t-2}', \ \dots \ , y_{t-p}']$, let U be the $T \times N$ matrix with rows u_t', $\Phi = [\Phi_0, \Phi_1, \ \dots \ , \Phi_p]'$, and $\Lambda = \text{diag}(\lambda_1, \ \dots \ , \lambda_T)$. The VAR in (8) can be expressed as $Y = X\Phi + U$ with likelihood function:

$$\begin{aligned} p(Y|\Phi, V, \Lambda) \propto |V|^{-0.5T} \exp\{-0.5 tr[V^{-1}(Y'\Lambda^{-1}Y - \Phi' X'\Lambda^{-1}Y \\ - Y'\Lambda^{-1}X\Phi + \Phi' X'\Lambda^{-1}X\Phi]\}. \end{aligned} \tag{11}$$

Note that the likelihood can be rearranged as follows:

$$\begin{aligned} p(Y|\Phi, V, \Lambda) \propto |V|^{-0.5k} \exp\{-0.5 tr[V^{-1}(\Phi - \hat{\Phi})'(X'\Lambda^{-1}X)(\Phi - \hat{\Phi})]\} \\ \times |V|^{-0.5(T-k)} \exp\{-0.5 tr[V^{-1}\hat{S}]\}, \end{aligned} \tag{12}$$

where $\hat{\Phi} = (X'\Lambda^{-1}X)^{-1}X'\Lambda^{-1}Y$ and $\hat{S} = (Y - X\hat{\Phi})'\Lambda^{-1}(Y - X\hat{\Phi})$ are the OLS estimator and the corresponding matrix of the sum of squared residuals. An inspection of the likelihood in (12) reveals that—conditional on the knowledge of Λ—it contains the kernels of an inverse Wishart distribution for V and a matricvariate normal distribution for Φ. Specifically, (12) is the density of a matricvariate normal-inverse Wishart (MNIW) distribution.[11]

[9] The estimates of the individual volatilities to which we refer here are based on univariate autoregressive models with stochastic volatility.

[10] For example, in 2005, Greenspan noticed that long-term interest rates became basically unresponsive to changes in short-term interest rates (see, among others, Bernanke, (2004)). Although including more than one factor would be a natural way to proceed, it would imply the loss of conjugacy, which in turn would make the Del Negro and Schorfheide (2004) approach unfeasible.

[11] A random variable Z has an MNIW distribution $Z \sim MNIW(M, P, S, \nu)$ when $Z|\Psi \sim MN(M, \Psi \otimes P)$ and $\Psi \sim IW(S, \nu)$.

3.3 | Priors

In the previous subsection, we have shown that the likelihood is MNIW. For this case, a conjugate prior for the coefficients Φ and V is viable. The general form of such a prior is $\Phi|V \sim N(\Phi_0, V \otimes \Omega_0)$, $V \sim IW(S_0, \nu_0)$, or compactly $\Phi, V \sim MNIW(\Phi_0, \Omega_0, S_0, \nu_0)$. The fact that the ATSM model described in Section 2 is linear and Gaussian ensures that the mean and variance of the state space (6) and (7) are sufficient statistics and therefore can be used to elicit the prior objects $\Phi_0, \Omega_0, S_0, \nu$.

In this paper, we adopt a hierarchical approach in which the prior moments $\Phi_0, \Omega_0, S_0, \nu_0$ are specified as functions of some hyperparameters: those in the set θ, which collects the coefficients of an underlying reference model, and the scalar γ, which measures the tightness with which one imposes such a reference model on the data. In the application, the reference model will be the JSZ model and θ the one in (5).

Following Del Negro and Schorfheide (2004), consider $T^* = \gamma T$ artificial observations from a restricted version of the VAR, which corresponds to an arbitrarily good approximation of some model of interest. Such a restricted version of the VAR has likelihood

$$p(Y|\Phi^*, V^*, \Lambda) \propto |V|^{-0.5T^*} \exp\{-0.5tr[\gamma T V^{-1}(\Gamma^*_{Y'Y}(\theta) - \Phi'\Gamma^*_{X'Y}(\theta) - \Gamma^*_{Y'X}(\theta)\Phi + \Phi'\Gamma^*_{X'X}(\theta)\Phi]\}, \tag{13}$$

where $\Gamma^*_{Y'Y}(\theta) = E_\theta[Y'\Lambda^{-1}Y]$, $\Gamma^*_{X'Y}(\theta) = E_\theta[X'\Lambda^{-1}Y]$, and $\Gamma_{X'X}{}^*(\theta) = E_\theta[X'\Lambda^{-1}X]$ are the moments of the (rescaled) data under the validity of the reference model. Conditionally on Λ, these moments can be computed from the state-space representation of the reference model using the autocovariance function (6 and 7) for any given parameterization θ in (5).

A standard way of interpreting a natural conjugate prior is as being the likelihood function from another sample. To see this point with reference to the density (13), imagine forming a dataset by adding the T^* artificial observations to T actual observations and estimate the VAR in (8) using such an augmented dataset. The likelihood of such a model would be the product of (11) and (13), and the likelihood moments would be a weighted average of the moments based on the artificial data and the actual data. The higher the ratio of artificial to actual observations ($\gamma = T^*/T$), the more weight is given to the restrictions implied by the reference model. Therefore, the artificial data act as prior information. Formally, the artificial data likelihood (13) can be rearranged as follows:

$$\begin{aligned} p(Y|\Phi, V, \Lambda) \propto |V|^{-0.5k} &\exp\{-0.5tr[V^{-1}(\Phi - \hat{\Phi}^*(\theta))'\Gamma^*_{X'X}(\theta)(\Phi - \hat{\Phi}^*(\theta))]\} \\ &\times |V|^{-0.5(T^*-k)} \exp\{-0.5tr[V^{-1}\hat{S}^*(\theta)]\}, \end{aligned} \tag{14}$$

where $\hat{\Phi}^*(\theta) = \Gamma^{*-1}_{X'X}(\theta)\Gamma^*_{X'Y}(\theta)$, $\hat{S}^*(\theta) = \gamma T(\Gamma^*_{Y'Y}(\theta) - \Gamma^*_{Y'X}(\theta)\Gamma^{*-1}_{X'X}(\theta)\Gamma^*_{X'Y}(\theta))$, and can be interpreted as a prior distribution for Φ and V:

$$\Phi, V|\theta, \gamma \sim MNIW\left(\hat{\Phi}^*(\theta), [\gamma T\Gamma^*_{X'X}(\theta)]^{-1}, \hat{S}^*(\theta), \gamma T - k\right). \tag{15}$$

The prior in (15) is hierarchical, that is, dependent on a second layer of coefficients: the hyperparameters in θ and γ. We use weakly informative priors for θ, implementing the belief that the first factor is a random walk, the second is stationary but very persistent, and the third is moderately persistent. For γ, we use a weakly informative Gaussian prior, truncated to satisfy the restriction $\gamma > (k + N)/T$, which is necessary for $p(\Phi, V|\theta, \gamma)$ to be proper. The prior mean for γ is centered on 1, which corresponds to giving a priori the same weight to the JSZ model and the unrestricted VAR. Finally, the prior specification is completed by adding a prior for the coefficients ϕ appearing in the volatility process.[12]

The joint prior distribution of all the coefficients of the model is

$$p(\Phi, V, \theta, \gamma, \phi) = p(\Phi, V|\theta, \gamma)p(\theta)p(\gamma)p(\phi). \tag{16}$$

More details on the distributions and moments will be contingent on the specific application at hand and will be spelled out in Section 4.2.

[12]Note that we do not need a prior on the first observation of the volatility process λ_1 as in Cogley and Sargent (2005), because in our setup, this value is set to 1 to identify the variance matrix of the u_t disturbances.

3.4 | Joint density of the full model

The full model is composed of the data Y, the latent states Λ, and the coefficients $\{\Phi, V, \theta, \gamma, \phi\}$. The joint density of data and states is the product of the likelihood (12) and the density of the log-normal states (10):

$$\begin{aligned} p(Y, \Lambda|\Phi, V, \theta, \gamma, \phi) &= p(Y|\Phi, V, \theta, \gamma, \phi, \Lambda) \times p(\Lambda|\Phi, V, \theta, \gamma, \phi) \\ &\propto p(Y|\Phi, V, \theta, \gamma, \Lambda) p(\Lambda|\phi), \end{aligned} \tag{17}$$

where the second line follows from omitting any redundant coefficients.

The joint density of data, coefficients, and states is obtained by multiplying (17) by the prior (16):

$$p(Y, \Phi, V, \theta, \gamma, \phi, \Lambda) = p(Y|\Phi, V, \theta, \gamma, \Lambda) p(\Lambda|\phi) \times p(\Phi, V|\theta, \gamma) p(\theta) p(\gamma) p(\phi). \tag{18}$$

3.5 | Conditional posterior distributions and MCMC sampler

Because $p(Y, \Phi, V, \theta, \gamma, \phi, \Lambda) = p(\Phi, V, \theta, \gamma, \phi, \Lambda|Y)p(Y)$, the density (18) is also the kernel of the joint posterior density of parameters and states. This does not correspond to any known distribution; hence, we simulate it via a MCMC algorithm:

1. Draw $\Phi, V, \theta, \gamma|Y, \Lambda$
   (a) Draw $\gamma, \theta|Y, \Lambda$(random walk Metropolis step)
   (b) Draw $\Phi, V|\theta, \gamma, Y, \Lambda$ (direct Monte Carlo step)
2. Draw $\Lambda|\Phi, V, \phi, Y$(independence Metropolis step)
3. Draw $\phi|\Phi, V, \Lambda, Y$

Conceptually, the algorithm is a Gibbs sampler, as each of the three steps draws from the conditional posterior distributions of the parameters and the states. Some of the steps are performed using a Metropolis step. Step 1 is performed via a random walk Metropolis step followed by direct Monte Carlo sampling, whereas Step 2 is performed via a sequence of Independence Metropolis steps. Note that this algorithm encompasses the one of Del Negro and Schorfheide (2004) as a special case.[13] This of course reflects the fact that the approach presented in this paper is a generalization of their approach to the case of a heteroskedastic VAR. We now turn to describing the steps in more detail.

3.5.1 | Conditional posteriors of hyperparameters and VAR coefficients (Step 1)

The density of interest can be factorized as $p(\Phi, V, \theta, \gamma|Y, \Lambda) \propto p(\Phi, V|\theta, \gamma, Y, \Lambda)p(\theta, \gamma|Y, \Lambda)$, and draws can be obtained by (1a) drawing from $\theta, \gamma|Y, \Lambda$ and (1b) drawing from $\Phi, V|\theta, \gamma, Y, \Lambda$.

In Step 1a, draws from $\theta, \gamma|Y, \Lambda$ can be further blocked into draws from $\gamma|\theta, Y, \Lambda$ and $\theta|\gamma, Y, \Lambda$. The conditional posterior densities $p(\gamma|\theta, Y, \Lambda)$ and $p(\theta|\gamma, Y, \Lambda)$ are proportional to the product of the priors $p(\gamma)$ and $p(\theta)$ and the marginal data density:

$$\begin{aligned} p(Y|\theta, \gamma, \Lambda) &= p(Y|\Phi, V, \Lambda) p(\Phi, V|\theta, \gamma, \Lambda) / p(\Phi, V|Y, \Lambda) \\ &= \frac{\left|\gamma T \Gamma^*_{X'X}(\theta) + X'\Lambda^{-1}X\right|^{-\frac{q}{2}} \left|\tilde{S}(\theta)\right|^{-\frac{(\gamma+1)T-k}{2}}}{\left|\gamma T \Gamma^*_{X'X}(\theta)\right|^{-\frac{q}{2}} \left|\tilde{S}^*(\theta)\right|^{-\frac{\gamma T-k}{2}}} \\ &\quad \times (2\Pi)^{\frac{-qT}{2}} \frac{2^{\frac{q((\gamma+1)T-k)}{2}} \Pi_{i=1}^{q} \Gamma[((\gamma+1)T-k+1-i)/2]}{2^{\frac{q(\gamma T-k)}{2}} \Pi_{i=1}^{q} \Gamma[(\gamma T-k+1-i)/2]}, \end{aligned} \tag{19}$$

where $\Gamma[\cdot]$ denotes the gamma function. The kernels $p(Y|\theta, \gamma, \Lambda)p(\gamma)$ and $p(Y|\theta, \gamma, \Lambda)p(\theta)$ are then used as target density in a random walk Metropolis step.

[13] In Del Negro and Schorfheide (2004), Steps 2 and 3 are absent (because there is no time variation in volatility) and Step 1a involves only θ(because the hyperparameter γ is fixed and not estimated).

In Step 1b, the conditional posterior density kernel of $p(\Phi, V|\theta, \gamma, Y, \Lambda)$ is given by the product of the likelihood (12) and the prior density (15), which gives

$$P(Y|\Phi, V, \Lambda) \propto |V|^{-0.5k} \exp\{-0.5tr[V^{-1}(\Phi - \tilde{\Phi}(\theta))'\Gamma^*_{X'X}(\theta)(\Phi - \tilde{\Phi}(\theta))]\} \times |V|^{-0.5(T+T^*-k)} \exp\{-0.5tr[V^{-1}\tilde{S}(\theta)]\}, \quad (20)$$

with $\tilde{\Phi}(\theta) = (\gamma T\Gamma^*_{X'X}(\theta) + X'\Lambda^{-1}X)^{-1}(\gamma T\Gamma^*_{X'Y}(\theta) + X'\Lambda^{-1}Y)$, $\tilde{S}(\theta) = [(\gamma T\Gamma^*_{Y'Y}(\theta) + Y'\Lambda^{-1}Y) - (\gamma T\Gamma^*_{Y'X}(\theta) + Y'\Lambda^{-1}X)(\gamma T\Gamma^*_{X'X}(\theta) + X'\Lambda^{-1}X)^{-1}(\gamma T\Gamma^*_{X'Y}(\theta) + X'\Lambda^{-1}Y)]$. The above is the kernel of an MNIW density:

$$\Phi, V|Y, \Lambda, V, \theta, \gamma \sim MNIW\left(\tilde{\Phi}(\theta), [\gamma T\Gamma^*_{X'X}(\theta) + X'\Lambda^{-1}X]^{-1}, \tilde{S}(\theta), (\gamma + 1)T - k\right). \quad (21)$$

When $\gamma \to 0$ the posterior mean of Φ approaches the *OLS* estimator. On the other hand, when $\gamma \to \infty$, the posterior mean of Φ approaches the prior mean $\Phi^*(\theta)$, that is, the value consistent with the JSZ model. Draws from (21) can be obtained by drawing from $V \sim IW(\tilde{S}(\theta), (\gamma + 1)T - k)$ and then from $\Phi|V \sim MN(\tilde{\Phi}(\theta), V \otimes [\gamma T\Gamma^*_{X'X}(\theta) + X'\Lambda^{-1}X]^{-1})$.

3.5.2 | Conditional posterior of latent states and their dynamics (Step 2 and Step 3)

Steps 2 and 3 of the algorithm produce draws from the volatility process Λ and its law of motion parameters ϕ, conditional on the VAR coefficients. Note that in these steps, conditioning on the hyperparameters θ and γ is redundant because under the knowledge of Φ and V, these hyperparameters do not yield any additional information.

To draw from the conditional posterior of Λ, we modify the approach of Cogley and Sargent (2005) to allow for a single stochastic volatility factor. Defining the orthogonalized residuals $w_t = (w_{1t}, \ldots, w_{Nt}) = V^{-1/2}u_t$, the kernel of $p(\Lambda|Y, \Phi, V, \phi)$ is given by

$$p(\Lambda|Y, \Phi, V, \theta, \gamma, \phi) \propto \prod_{t=2}^{T} p(\lambda_t|\lambda_{t-1}, \lambda_{t+1}, \phi, w_t). \quad (22)$$

By choosing an appropriate proposal density, this kernel can be used as a basis for an independence Metropolis step with acceptance probability

$$a = \min\left(\frac{\lambda_t^{*-N\times 0.5}\prod_{i=1}^{N}\exp(-0.5w_{it}^2/\lambda_t^*)}{\lambda_t^{-N\times 0.5}\prod_{i=1}^{N}\exp(-0.5w_{it}^2/\lambda_t)}, 1\right). \quad (23)$$

Note that this differs from Cogley and Sargent (2005), as in their case, the volatility process λ_{it} for each variable i is drawn separately conditional on the remaining $N - 1$ terms, which means that $N - 1$ elements in the products $\Pi_{i=1}^{N}\exp(-0.5w_{it}^2/\lambda_t^*)$ and $\Pi_{i=1}^{N}\exp(-0.5w_{it}^2/\lambda_t)$ would cancel out. Details on the derivation are provided in Appendix S1, Section C.

Finally, Step 3 is straightforward because (10) is a linear regression model, and conditionally on Λ, the conditional posterior distributions of ϕ are readily available.

3.6 | Homoskedastic version

In our forecasting exercise, we will also consider a homoskedastic version of the model. This is simply obtained by setting $\lambda_t = 1$ for all t, yielding

$$y_t = \Phi_0 + \Phi_1 y_{t-1} + \ldots + \Phi_p y_{t-p} + \epsilon_t; \quad \epsilon_t \sim N(0, V). \quad (24)$$

This model is nested in the more general JSZ-VAR-CSV shown in Equations (8)–(10), the only difference being that the volatility is assumed to be constant over time, and therefore, the parameters ϕ and the volatility λ_t drop out of the analysis. We label this version JSZ-VAR. For estimation, the JSZ-VAR requires a simplified algorithm:

1. Draw $\Phi, V, \theta, \gamma|Y, \Lambda$
   (a) Draw $\gamma, \theta|Y, \Lambda$(random walk Metropolis step)
   (b) Draw $\Phi, V|\theta, \gamma, Y, \Lambda$ (direct Monte Carlo step)

This algorithm coincides with Step 1 of Algorithm 1, and it is very similar to the one of Del Negro and Schorfheide (2004), the only difference being that in their paper the coefficient γ is not estimated, so Step 1a only involves θ.

3.7 | ZLB treatment and shadow rates

Since the 2007–2008 financial crisis, interest rates decreased to a point that a zero lower bound (ZLB) cannot be ignored. A natural way to include the ZLB restriction would be to directly modify the term structure model in order to implement the constraint and allow for estimation of shadow rates. This approach has been pursued in the shadow rate term structure models of studies such as Krippner (2013) and Wu and Xia (2016).

In our approach, the imposition of a ZLB in the reference model would break down Gaussianity, which in turn would imply that the moments of the state-space system would not be sufficient statistics and could not be used to form a prior on the VAR coefficients. More importantly, introducing a ZLB in the reference model would not be sufficient to ensure that the ZLB is satisfied in the posterior estimates nor would it allow filtering out a shadow rate. In order to effectively implement the ZLB, we need to impose it on the entire VAR system.

To do so, we follow the approach of Johannsen and Mertens (2021) and frame the issue as a censored data problem. The observed yield of maturity τ is defined as $y_t^\tau = \max(s_t^\tau, 0)$, where s_t^τ is a shadow yield that can be either positive or negative. The shadow yield is the—possibly negative—yield that one would observe if the ZLB constraint did not exist. In normal periods $y_t^\tau = s_t^\tau > 0$, but in periods in which the ZLB is binding, $y_t^\tau = 0$ and s_t^τ is negative and unobservable.

With this extension, the MCMC sampler needs an additional step involving a draw from the conditional posterior of the shadow yields s_t^τ. This draw of s_t^τ is then used as data in place of y_t^τ in the remaining steps of the algorithm, in any time period t and for any maturity τ in which the yield y_t^τ hits the lower bound. More details can be found in Johannsen and Mertens (2021), who show how to filter out the shadow yields in a general VAR.

In the forecasting exercise, the shadow yields s_t^τ are used instead of the actual yields y_t^τ to produce forecasts of the shadow yields s_{t+h}^τ. Then, we impose the ZLB by truncating all of the forecast paths going below zero, that is, we set $y_{t+h}^\tau = \max(s_{t+h}^\tau, 0)$. This strategy is implemented for all yields at all forecast origins, and it truncates the entire predictive density of the model(s) in such a way that the ZLB is satisfied, while using all of the additional information contained in the shadow rates.

4 | EMPIRICAL APPLICATION

We now turn to the application of our method using US data. All of the results in the paper are based on four independent MCMC chains.[14] Table S1 shows that the algorithm has good mixing properties and has achieved convergence.

4.1 | Data

Data are zero-coupon yields, at monthly frequency, for maturities 1 and 3 months and 1, 2, 3, 4, 5, 7, and 10 years. We obtained the yields from maturities of 1 month through 5 years from the US Treasuries Daily and Monthly database of the Center for Research in Security Prices (CRSP), the University of Chicago Booth School of Business. We took the 7- and 10-year yields from the Gürkaynak et al. (2007) data published by the Federal Reserve Board of Governors Board. Our sample extends from January 1985 through December 2018. The data are displayed in Figure 1. We estimate all of our VAR specifications using three lags, chosen via the Bayesian information criterion.

4.2 | Specifics on priors

Recall the prior in (16):

$$p(\Phi, V, \theta, \gamma, \phi) = p(\Phi, V|\theta, \gamma)p(\theta)p(\gamma)p(\phi). \tag{25}$$

The priors on the VAR coefficient matrices $p(\Phi, V|\theta, \gamma)$ are set up hierarchically, using the MNIW distribution set out in (15). Therefore, we only need to specify priors for γ, θ, ϕ.

For the parameter γ, which is measuring the degree with which JSZ-consistent moments are imposed on the VAR, we set a normal prior centered on 1, with a standard deviation of 0.25. We truncate the posterior draws by requiring them to be above $(k + N)/T$, as this is the minimum value necessary for the priors on V and Φ to be proper. The prior mean of

[14]Each chain is composed of 15,000 draws. We eliminate the first 2500 as burn-in, and we perform skip sampling, retaining each 25th draw, for a total of 500 clean draws per chain. This provides 2000 clean draws in total. We initialize each chain using the posterior mode of the model, conditional on a maximum likelihood estimate of the common volatility factor, plus a random error, so that each chain has a different initial condition.

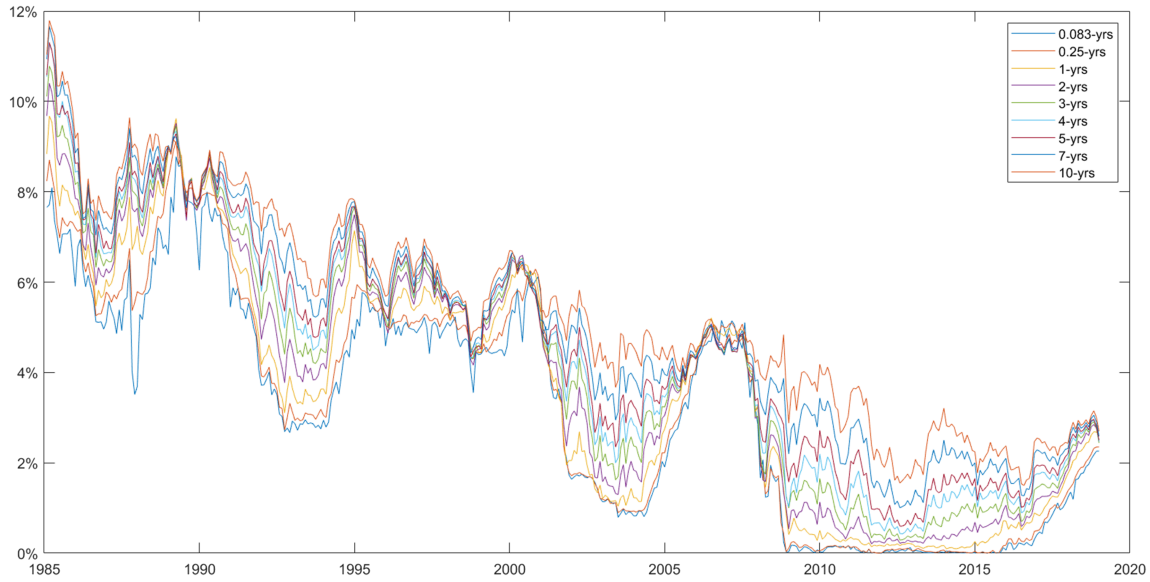

FIGURE 1 US zero-coupon yields, January 1985 to December 2018 [Colour figure can be viewed at wileyonlinelibrary.com]

1 reflects the belief that the JSZ model and an unrestricted VAR are equally likely descriptions of the data. The standard deviation of 0.25 is rather large and implies that our prior is only weakly informative.[15]

For the JSZ structural parameters θ, we set either a flat or a weakly informative prior. In particular, for the three coefficients in $\lambda^{\mathbb{Q}}$, we set a normal prior $\lambda_1^{\mathbb{Q}} \sim N(-0.002, 0.001^2)$, $\lambda_2^{\mathbb{Q}} \sim N(-0.02, 0.01^2)$, $\lambda_3^{\mathbb{Q}} \sim N(-0.2, 0.1^2)$. These prior means imply that under the $\mathbb{Q}$ measure, the first factor is virtually a random walk (it features an autoregressive coefficient of 0.998), the second is stationary but very persistent (with an autoregressive coefficient of 0.980), and the third factor is moderately persistent (with an autoregressive coefficient of 0.800). All draws of $\lambda^{\mathbb{Q}}$ implying nonstationary behavior are discarded, as well as all those for which the condition $\lambda_1^{\mathbb{Q}} > \lambda_2^{\mathbb{Q}} > \lambda_3^{\mathbb{Q}}$ does not hold.[16] For the coefficients Σ_P, we set a normal prior centered on the VAR of the observable factors P_t^o.[17] We set the standard deviations to half of the prior means, which ensures that a 95% credible interval for each coefficient is marginally above 0. For the remaining coefficients $k_\infty^{\mathbb{Q}}$ and Σ_y, we set a uninformative flat prior. Finally, for the parameters governing the dynamics of the volatility factor ϕ, we set $\phi_0 \sim N(0, 0.025)$, $\phi_1 \sim N(0.96, 0.025)$, and $\phi_2 \sim IG(3 \cdot 0.05, 3)$.

The priors described above are only weakly informative, and the resulting posterior estimates are a fair amount away from them. To check for robustness, we have also computed results for a more diffuse prior in which the standard deviation of γ is set to 1 and the prior on θ is flat. This changed somewhat the estimates of γ, making them higher. This is driven by the fact that the flat prior on θ is imposing the JSZ model more blandly than before, and therefore, the overall level of misspecification of the model decreases, which makes γ increase. However, the overall mass of the posterior of γ is unchanged, and the posterior means of all of the remaining parameters were very similar.

4.3 | In-sample results

We start with in-sample estimation of the baseline model, the JSZ-VAR-CSV given in (8)–(10), with the JSZ prior described in Section 3.3.

Table 1 contains estimates of the hyperparameters θ shown in (5). These are the coefficients of the underlying JSZ model. Under the heteroskedastic JSZ-VAR-CSV model, the estimated posterior means of the coefficients $\lambda_1^{\mathbb{Q}}, \lambda_2^{\mathbb{Q}}, \lambda_3^{\mathbb{Q}}$ are −0.00307, −0.03533, and −0.07549, respectively.[18] These are broadly in line with the values −0.00245, −0.0472, and

[15]The values of γ range from 0.3 to 1.2 throughout the recursive samples.
[16]This is required because $\lambda_1^Q, \lambda_2^Q, \lambda_3^Q$ are ordered eigenvalues of the matrix K_{1S}^Q, which is in Jordan form.
[17]In principle, one should not use likelihood information to calibrate the prior, but doing so for error variances using an auxiliary model is standard practice; see, for example, Doan et al. (1984), Litterman (1986), Sims (1993), Robertson and Tallman (1999), Sims and Zha (1998), Kadiyala and Karlsson (1997), Banbura et al. (2010), Koop (2013), and Carriero et al. (2015).
[18]To improve readability, Table 1 reports results for $100 \times \theta$.

TABLE 1 Structural coefficient estimates

Coefficient (×100)	JSZ-VAR-CSV 1985:1-2018:12	JSZ-VAR-CSV 1985:1-2007:12	JSZ-VAR (homoskedastic) 1985:1-2018:12	JSZ-VAR (homoskedastic) 1985:1-2007:12
λ_1^Q	−0.307	−0.234	−0.265	−0.243
	0.087	0.077	0.082	0.078
λ_2^Q	−3.533	−2.829	−3.227	−2.719
	0.582	0.353	0.514	0.374
λ_3^Q	−7.549	−10.825	−7.639	−10.707
	1.43	1.493	1.620	1.826
κ^Q	0.037	0.034	0.015	0.036
	0.010	0.006	0.008	0.006
$\Sigma_{P(1,1)}$	0.637	0.603	0.600	0.665
	0.074	0.081	0.042	0.048
$\Sigma_{P(2,1)}$	0.225	−0.180	0.147	−0.148
	0.034	0.030	0.028	0.026
$\Sigma_{P(2,2)}$	0.263	0.243	0.257	0.274
	0.034	0.036	0.020	0.022
$\Sigma_{P(3,1)}$	−0.086	−0.078	−0.068	−0.065
	0.015	0.014	0.014	0.012
$\Sigma_{P(3,2)}$	−0.025	0.049	−0.038	0.059
	0.016	0.015	0.014	0.014
$\Sigma_{P(3,3)}$	0.139	0.099	0.123	0.106
	0.020	0.015	0.011	0.010
$\Sigma_{y(1,1)}$	0.172	0.184	0.188	0.216
	0.024	0.027	0.015	0.017
$\Sigma_{y(2,2)}$	0.047	0.047	0.048	0.043
	0.015	0.015	0.014	0.018
$\Sigma_{y(3,3)}$	0.060	0.056	0.055	0.059
	0.008	0.009	0.005	0.005
$\Sigma_{y(4,4)}$	0.030	0.027	0.030	0.031
	0.004	0.005	0.003	0.003
$\Sigma_{y(5,5)}$	0.021	0.020	0.021	0.024
	0.004	0.004	0.003	0.003
$\Sigma_{y(6,6)}$	0.022	0.019	0.024	0.022
	0.003	0.003	0.002	0.002
$\Sigma_{y(7,7)}$	0.027	0.026	0.028	0.031
	0.004	0.004	0.002	0.003
$\Sigma_{y(8,8)}$	0.004	0.004	0.004	0.005
	0.003	0.003	0.003	0.003
$\Sigma_{y(9,9)}$	0.026	0.022	0.024	0.025
	0.004	0.004	0.003	0.003

Note: Estimates of the structural coefficients θ of the reference JSZ model. The entries are posterior means and standard deviations (in smaller size) computed from the MCMC output. To improve readability, the table reports results for $100 \times \theta$.

−0.0739, reported by JSZ in their RKF specification.[19] Our estimate of $k_\infty^{\mathbb{Q}}$ is 0.00037, which corresponds to a value for the long-run mean of the short-term rate under the risk-neutral measure of $-k_\infty^{\mathbb{Q}}/\lambda_1^{\mathbb{Q}} = 11.9$, which is somewhat off the value reported by JSZ in their RKF specification (8.45) but very close to the value of 11.2 they obtain with some other specifications (RCMT and RCMT_1). The tightness hyperparameter γ (not in the table) is estimated at 0.358, signaling that the optimal ratio between artificial and actual observations is about 36%, which means that the data are given about twice the weight of the prior.

Table 1 also reports results based on the homoskedastic JSZ-VAR specification, which are very similar. In this case, the hyperparameter γ (not in the table) is estimated at 0.314, which is slightly lower than in the heteroskedastic case. With

[19] JSZ present results for various specifications. The RKF specification is based on a model with all yields measured with errors and is the closest to our reference model. Results for the other specifications are all in the same ballpark. Of course, our dataset is different, both in the time series and in the cross-sectional dimension.

this model, more weight is given to the data and less to the prior (compared with the JSZ-VAR-CSV specification), which is a sign that the JSZ prior based on the nonrescaled data has a slightly higher degree of misspecification.

Finally, to ascertain the stability of the estimates before and after the financial crisis, the table reports results based on a sample ending in December 2007. The estimated coefficients are broadly stable, with a few exceptions, notably the coefficients $\Sigma_{P(2,1)}$ and $\Sigma_{P(3,2)}$ changing sign, which points to a change in the direction of correlation among the factors. The coefficient $\lambda_3^{\mathbb{Q}}$ signals an increase in the persistence of the third factor in the second half of the sample, but the change is small relative to the precision of the estimate so it might be insignificant.

Turning to the VAR model estimates, Figure 2 displays the posterior distribution of the CSV factor λ_t, and Figure 3 displays the implied time series of the stochastic volatilities for each of the yields in the VAR, that is, the diagonal of the variance matrix of the disturbances u_t (see Equation 9). As is well known, there have been periods of high and of low volatilities throughout the sample under examination, and this is captured in our estimates.

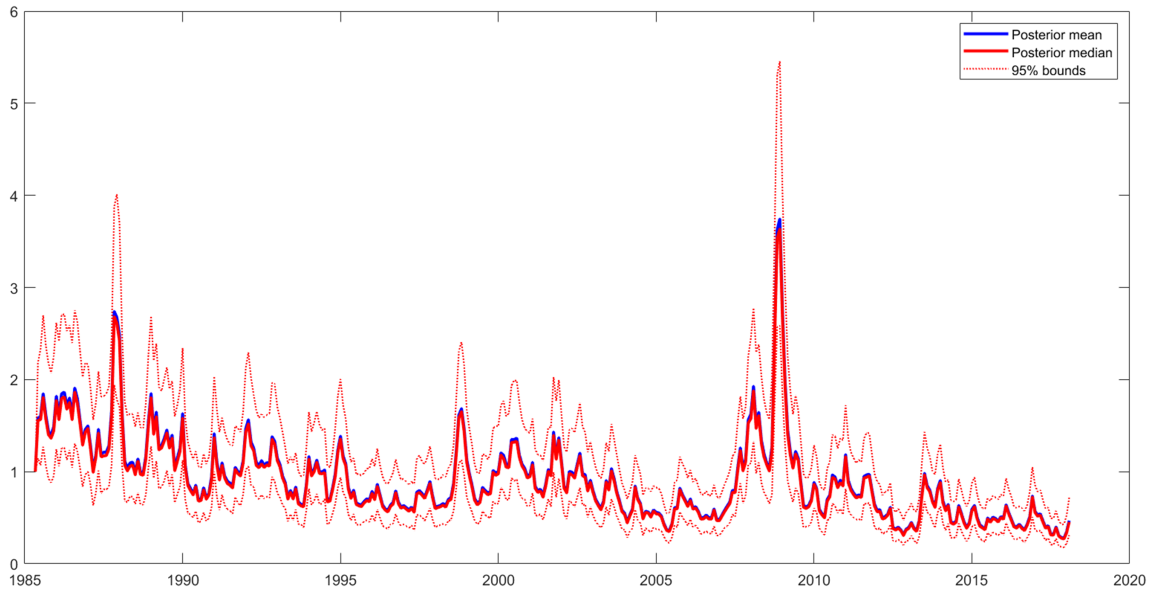

FIGURE 2 Posterior distribution of the common stochastic volatility process λ_t [Colour figure can be viewed at wileyonlinelibrary.com]

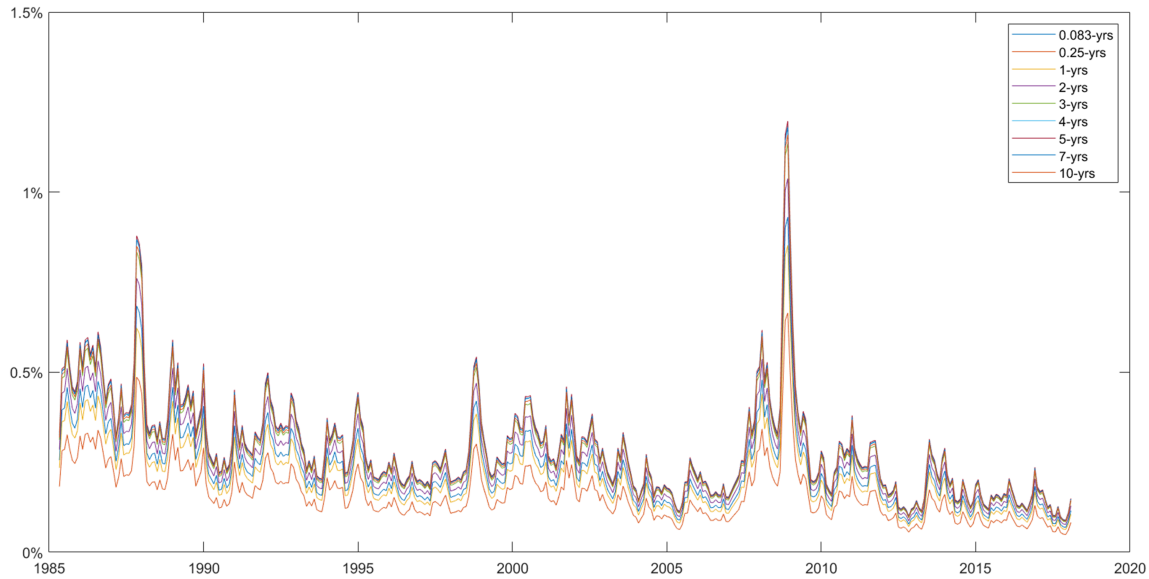

FIGURE 3 Volatilities for each yield (posterior medians) [Colour figure can be viewed at wileyonlinelibrary.com]

4.4 | Forecasting exercise

4.4.1 | Models

We will compare the out-of-sample forecasting performance of the baseline JSZ-VAR-CSV model with several alternatives, some taken from the literature and others variations on the JSZ-VAR-CSV specification. As the benchmark to which all forecasts are compared, we use a simple random walk forecast, which is simple but nonetheless has proven extremely strong in forecasting the term structure of government bond yields.[20] Another previously developed model we will consider is a fully fledged ATSM with time-varying volatility. Specifically, we use the dynamic Nelson–Siegel specification of Diebold and Li (2006) with the addition of time-varying volatility in the factors, as in Hautsch and Ou (2012).[21,22]

Our baseline model is the JSZ-VAR-CSV model, featuring three lags, time variation in volatility, and shrinkage toward the JSZ representation, which entails both the assumption that yields follow a factor model and some restrictions on the loadings. To assess the contribution of each of these ingredients to forecast accuracy, we consider alternative models in which some of these features are removed.

First, in order to assess whether any forecast gains come from shrinkage per se rather than from the specific direction of shrinkage, we include in the comparison an alternative VAR-CSV that shrinks toward some other direction than the JSZ model. Specifically, we set the prior mean and variances of a Minnesota-style prior, in which the yields follow univariate random walks.[23] For implementation, we compute the relevant moment matrices based on this alternative prior, and we use them in (15). Besides the use of these alternative prior moment matrices, all of the remaining characteristics of the model are left unchanged and the model is estimated using the same MCMC sampler as the baseline. In particular, the overall tightness on this prior is optimally chosen by estimating the parameter γ via a Metropolis step. We label this model BVAR-CSV in the tables.

Second, in order to ascertain the role of time variation in volatility, we consider the homoskedastic version of the model described in Section 3.6, labeled JSZ-VAR. This version features shrinkage toward the JSZ restrictions, but the volatilities are kept constant over time. Moreover, although our baseline model does feature time variation in the volatilities of the yields, it is a Bayesian VAR with a prior based on an ATMS and not itself simply an ATSM.

Third, a relevant difference between our baseline model and a typical term structure model is the lag length. Term structure models typically use only one lag, but research has shown that yields are not necessarily first-order Markov (Joslin et al., 2013a, 2013b). For this reason, we include in the comparison a version of the baseline model with only one lag, in order to gauge the effect that the richer dynamics have on the forecasts. We label this version of the model JSZ-VAR-CSV (one lag).

Finally, as we have discussed, the JSZ representation entails both the imposition of a factor model for the yields and a set of restrictions on the factor loadings. We want to disentangle the role played by these two elements in the forecasting outcome. In order to do so, we have re-estimated the model using a reference prior that only implements the factor structure for the yields, without also imposing the restrictions on the loadings. This can be done by simply concentrating out, via an OLS regression on the principal component of the yields, the coefficients in the vector A_p and the matrix B_p appearing in (2), which implies that these coefficients are no longer obeying the Riccati equations, and therefore, they do not necessarily reflect the absence of arbitrage.[24] We label this model F-VAR-CSV.

[20] Term structure models have a hard time improving on the accuracy of a simple random walk forecast, as documented in several studies, including Duffee (2002), Diebold and Li (2006), Christensen et al. (2011), and Carriero et al. (2012).

[21] Although extending the JSZ model to allow for time-varying volatility is possible (and has been done by Creal & Wu, (2015)), the complications inherent in its estimation are such that it is prohibitive to estimate such a model repeatedly for an out-of-sample forecasting exercise the size of ours. Moreover, several contributions have shown that models with unspanned stochastic volatility such as Hautsch and Ou (2012) are preferable as they tend to fit the yields better than models with spanned stochastic volatility.

[22] In line with the JSZ specification, we depart from Hautsch and Ou (2012) in using a VAR (not AR) for factor dynamics.

[23] The Minnesota-style prior we implement is the same as that in Kadiyala and Karlsson (1997), augmented with the "sum of coefficients" and "dummy initial observation" priors proposed in Doan et al. (1984) and Sims (1993), with the hyperparameter choice of Sims and Zha (1998). Both priors are in line with the belief that macroeconomic data typically feature unit roots, and the latter prior favors cointegration. This prior is similar to that of Sims and Zha (1998), with the subtle difference that in the original implementation the prior is elicited on the coefficients of the structural representation of the VAR rather than on the reduced form as it is here. This prior has been widely used in the literature, which documented its competitiveness in forecasting macroeconomic data; see, for examples, Carriero et al. (2015), Giannone et al. (2015), Robertson and Tallman (1999), and Waggoner and Zha (1999).

[24] We could also estimate these coefficients within the MCMC algorithm. We choose the simpler route because our goal is simply to establish the relative importance of the cross-equation restrictions.

4.4.2 | Design

We perform an out-of-sample forecasting exercise. We start with an estimation window ranging from January 1985 to December 1994, we estimate the model, and we produce forecasts for the period January 1995 to December 1995 (i.e., up to 12 steps ahead). Then, we add one data point to the sample, namely, January 1996, and we re-estimate the model and again produce forecasts up to 12 steps ahead. We proceed in this way until we obtain forecasts for the period January 2018 to December 2018.

We obtain forecast distributions by sampling as appropriate from the posterior distributions of the considered models. For example, in the case of the JSZ-VAR-CSV model, for each set of draws of parameters, we (1) simulate volatility time paths over the forecast interval using the AR(1) structure of log volatility, (2) draw shocks to each variable over the forecast interval with variances equal to the draw of V_{t+h}, and (3) use the VAR structure of the model to obtain paths of each variable. For all of the models described above, we impose a ZLB by using the method of Johannsen and Mertens (2021) briefly described in Section 3.7. We form point forecasts as means of the draws of simulated forecasts and density forecasts from the simulated distribution of forecasts. Conditional on the model, the posterior distribution reflects all sources of uncertainty (latent states, parameters, hyperparameters, and shocks over the forecast interval). For the random walk, point forecasts are set to the value of the yields in the previous period. Density forecasts are produced by simulating yields over the forecast interval using a random walk specification with innovations variance equal to the variance of changes in yields over the estimation sample.

4.4.3 | Forecast evaluation

We evaluate both point and density forecasts of the examined models. For point forecasts, we evaluate our results in terms of root mean squared forecast errors (RMSFEs) for a given model. Let $\hat{y}^{(i)}_{t+h|t}(M)$ denote the h-step-ahead point forecast (mean of the predictive density) made by model M at time t, for the ith yield. The RMSFE made by model M in forecasting the ith yield at horizon h is $\text{RMSFE}^{M}_{i,h} = \sqrt{P^{-1}\sum\left(\hat{y}^{(i)}_{t+h|t}(M) - y^{(i)}_{t+h}\right)^2}$, where the sum is computed over all of the P forecasts.

To provide a rough gauge of whether the RMSFE ratios are significantly different from 1, we use the Diebold and Mariano (1995) t-statistic for equal MSE, applied to the forecast of each model relative to the benchmark. Our use of the Diebold–Mariano test with forecasts that are, in some cases, nested is a deliberate choice. Monte Carlo evidence in Clark and McCracken ((2011), (2015)) indicates that, with nested models, the Diebold–Mariano test compared against normal critical values can be viewed as a somewhat conservative (conservative in the sense of tending to have size modestly below nominal size) test for equal accuracy in the finite sample. As our proposed model can be seen as nesting the benchmarks we will compare it against, we treat the tests as one sided and only reject the benchmark in favor of the null (i.e., we do not consider rejections of the alternative model in favor of the benchmark).

The overall accuracy of the density forecasts can be measured with average log predictive density scores, motivated and described in such sources as Geweke and Amisano (2010), given by $\text{SCORE}^{M}_{i,h} = P^{-1}\sum \log p(y^{(i)}_{t+h}|y_t, M)$, where the sum is computed over all of the P forecasts and where $\log p(y^{(i)}_{t+h}|y_t, M)$ is the log predictive score obtained by model M, at forecast origin t, when making the h-step-ahead forecast of the ith yield. We compute the log predictive scores using the quadratic approximation of Adolfson et al. (2007).[25]

To provide a rough gauge of the statistical significance of differences, we use the Amisano and Giacomini (2007) t-test of equal means, applied to the log score for each model relative to the benchmark random walk forecast. We view the tests as a rough gauge because, with nested models, the asymptotic validity of the Amisano and Giacomini (2007) test requires that, as forecasting moves forward in time, the models be estimated with a rolling, rather than an expanding, sample of data. As our proposed model can be seen as nesting the benchmarks that we will compare it against, we treat the tests as one sided and only reject the benchmark in favor of the null (i.e., we do not consider rejections of the alternative model in favor of the benchmark).

4.5 | Out-of-sample results

We now turn to the empirical evidence on the forecasting performance of the proposed model. The forecast evaluation period goes from January 1995 to December 2018, and forecasts are produced as described in Section 4.4.2.

[25]We obtained very similar results using a kernel density approximation, except that these ran into numerical problems in a couple of forecast origins.

Tables 2 and 3 present results for point and density forecasts, respectively. In the tables, the first panel contains the RMSFEs and SCOREs obtained by the random walk forecasts of the yields (the units are basis points). The remaining panels display—for the remaining models—the RMSFEs *ratios* and SCOREs *differences* relative to the random walk;

TABLE 2 Evaluation of point forecasts: Sample 1995:2018

	Maturity								
Step ahead	**0.083 year**	**0.25 year**	**1 year**	**2 year**	**3 years**	**4 years**	**5 years**	**7 years**	**10 years**
	RW point forecasting performance								
1	27.648	20.036	21.165	24.470	26.233	27.695	27.426	27.323	26.827
2	38.344	31.311	33.485	37.693	39.504	41.003	40.484	39.889	38.841
3	47.200	42.635	45.029	48.417	49.363	50.123	49.182	46.834	44.608
6	74.254	72.084	72.649	72.540	71.143	70.866	69.667	65.050	61.360
12	125.407	126.882	120.670	109.451	99.858	94.040	89.856	81.825	76.100
	DNS-SV versus random walk								
1	1.068	1.141	1.094	1.080	1.109	1.155	1.185	1.045	1.094
2	1.072	1.138	1.074	1.116	1.133	1.152	1.174	1.066	1.021
3	1.111	1.141	1.094	1.142	1.161	1.174	1.195	1.104	1.025
6	1.140	1.150	1.146	1.207	1.236	1.239	1.249	1.181	1.077
12	1.115	1.103	1.140	1.252	1.336	1.377	1.400	1.359	1.241
	BVAR-CSV versus random walk								
1	0.941***	0.930***	0.983	1.004	1.008	1.008	1.007	1.001	0.998
2	0.922***	0.915***	0.995	1.029	1.027	1.023	1.019	1.010	1.002
3	0.937**	0.948*	1.026	1.047	1.041	1.032	1.024	1.012	1.002
6	0.965	0.997	1.093	1.112	1.097	1.078	1.062	1.032	1.007
12	0.998	1.030	1.137	1.180	1.183	1.168	1.149	1.090	1.029
	JSZ-VAR versus random walk								
1	0.852***	0.909***	1.008	1.058	1.040	1.045	1.043	1.039	1.052
2	0.816***	0.867**	0.991	1.081	1.066	1.065	1.062	1.054	1.059
3	0.835**	0.887*	0.995	1.080	1.072	1.065	1.062	1.055	1.055
6	0.839*	0.901	1.011	1.073	1.074	1.059	1.043	1.031	1.025
12	0.844	0.885	0.973	1.043	1.073	1.073	1.056	1.043	1.031
	JSZ-VAR-CSV versus random walk								
1	0.840***	0.967	1.001	1.046	1.036	1.046	1.039	1.028	1.037
2	0.825***	0.945	1.005	1.069	1.053	1.051	1.044	1.028	1.029
3	0.865**	0.958	1.005	1.062	1.051	1.043	1.040	1.026	1.024
6	0.903	0.972	1.039	1.080	1.066	1.049	1.032	1.012	1.005
12	0.911	0.947	1.014	1.072	1.084	1.079	1.058	1.032	1.014
	JSZ-VAR-CSV (one lag) versus random walk								
1	0.829***	0.984	1.018	1.061	1.037	1.043	1.041	1.028	1.032
2	0.831***	0.962	1.016	1.074	1.052	1.049	1.042	1.020	1.016
3	0.867**	0.971	1.023	1.083	1.067	1.060	1.049	1.022	1.010
6	0.904	0.967	1.037	1.087	1.077	1.060	1.039	1.010	0.993
12	0.921	0.951	1.015	1.074	1.085	1.074	1.045	1.002	0.968
	F-VAR-CSV versus random walk								
1	0.845***	0.925**	0.986	1.039	1.037	1.042	1.040	1.032	1.042
2	0.805***	0.874**	0.963	1.053	1.049	1.046	1.043	1.034	1.036
3	0.818***	0.881*	0.958	1.043	1.045	1.039	1.039	1.033	1.034
6	0.840*	0.899	0.988	1.056	1.056	1.048	1.036	1.026	1.023
12	0.859	0.892	0.973	1.036	1.054	1.057	1.042	1.026	1.013

Note: The first panel contains the RMSFEs obtained by using the random walk forecasts, with units in basis points. The remaining panels display the relative RMSFEs of the competing models relative to the random walk. A figure below 1 in the relative RMSFEs signals a model that is outperforming the random walk benchmark. Gains in accuracy that are statistically different from zero are denoted by asterisks, evaluated using the Diebold and Mariano (1995) t-statistics computed with a serial correlation-robust variance, using a rectangular kernel, $h - 1$ lags (h denotes the forecast horizon), and the small-sample adjustment of Harvey et al. (1997).

* Significance levels of 10%.

** Significance levels of 5%.

*** Significance levels of 1%.

TABLE 3 Evaluation of density forecasts: Sample 1995:2018

Step ahead	Maturity								
	0.083 year	0.25 year	1 year	2 years	3 years	4 years	5 years	7 years	10 years
	RW density forecasting performance								
1	−4.857	−4.470	−4.552	−4.662	−4.718	−4.778	−4.759	−4.746	−4.729
2	−5.184	−4.884	−4.962	−5.059	−5.107	−5.146	−5.128	−5.113	−5.099
3	−5.391	−5.181	−5.235	−5.300	−5.328	−5.344	−5.331	−5.283	−5.252
6	−5.800	−5.737	−5.698	−5.690	−5.691	−5.690	−5.669	−5.600	−5.562
12	−6.259	−6.476	−6.230	−6.100	−6.033	−5.992	−5.945	−5.884	−5.826
	DNS-SV versus random walk								
1	0.061	−0.072	0.013	−0.033	−0.106	−0.205	−0.334	−0.191	−0.285
2	0.068	−0.077	−0.015	−0.133	−0.211	−0.316	−0.444	−0.332	−0.251
3	0.028	−0.088	−0.070	−0.205	−0.288	−0.390	−0.513	−0.405	−0.245
6	−0.114	−0.155	−0.266	−0.431	−0.503	−0.588	−0.705	−0.572	−0.319
12	−0.403	−0.230	−0.565	−0.810	−0.919	−0.996	−1.082	−0.874	−0.525
	BVAR-CSV versus random walk								
1	0.349***	0.250***	0.199***	0.083***	0.027	0.018	−0.012	−0.020	−0.008
2	0.322***	0.219***	0.148***	0.039	0.006	−0.007	−0.035	−0.056	−0.055
3	0.284***	0.160***	0.091**	0.017	−0.002	−0.025	−0.034	−0.069	−0.073
6	0.187***	0.097	0.010	−0.018	−0.006	−0.001	−0.012	−0.017	0.004
12	−0.013	0.050	−0.107	−0.139	−0.119	−0.093	−0.097	−0.041	−0.000
	JSZ-VAR versus random walk								
1	0.301***	0.210***	0.096***	−0.006	−0.017	−0.004	−0.015	−0.012	−0.011
2	0.367***	0.247***	0.099***	−0.032	−0.044	−0.041	−0.044	−0.032	−0.016
3	0.365***	0.244***	0.101**	−0.027	−0.046	−0.050	−0.044	−0.042	−0.024
6	0.367***	0.281**	0.101	−0.013	−0.028	−0.031	−0.030	−0.034	−0.013
12	0.314***	0.473*	0.163	0.029	0.004	−0.003	−0.009	0.001	0.003
	JSZ-VAR-CSV versus random walk								
1	0.448***	0.301***	0.202***	0.059**	0.009	0.015	−0.019	−0.022	−0.011
2	0.453***	0.291***	0.185***	0.042	0.012	0.008	−0.020	−0.032	−0.032
3	0.405***	0.250***	0.155***	0.036	0.012	−0.003	−0.015	−0.046	−0.058
6	0.305***	0.233*	0.106	0.023	0.019	0.017	0.008	−0.013	−0.011
12	0.192**	0.375	0.116	0.008	−0.008	−0.005	−0.017	−0.011	−0.014
	JSZ-VAR-CSV (one lag) versus random walk								
1	0.441***	0.265***	0.191***	0.063**	0.017	0.016	−0.015	−0.017	−0.002
2	0.441***	0.256***	0.187***	0.045	0.019	0.007	−0.019	−0.030	−0.025
3	0.381***	0.200**	0.152***	0.030	0.006	−0.016	−0.021	−0.040	−0.039
6	0.272**	0.206*	0.108	0.024	0.019	0.017	0.012	0.002	0.014
12	0.118	0.307	0.083	−0.009	−0.017	−0.004	−0.007	0.010	0.010
	F-VAR-CSV versus random walk								
1	0.441***	0.338***	0.195***	0.050*	−0.001	0.004	−0.028	−0.027	−0.011
2	0.479***	0.348***	0.191***	0.038	0.006	0.001	−0.027	−0.039	−0.038
3	0.458***	0.319***	0.167***	0.036	0.009	−0.007	−0.017	−0.046	−0.055
6	0.360***	0.285**	0.108	0.018	0.012	0.008	0.000	−0.020	−0.016
12	0.219***	0.401	0.116	0.012	−0.002	−0.001	−0.012	−0.007	−0.010

Note: The first panel contains the average SCOREs obtained by using the random walk forecasts. The remaining panels display the differences in SCOREs of the competing models relative to the random walk. A figure above 0 in the SCORE differences signals that a model is outperforming the random walk benchmark. As the SCOREs are measured in logs, a score difference of, for example, 0.05 signals a 5% gain in terms of density forecast accuracy. Gains in accuracy that are statistically different from zero are denoted by asterisks, evaluated using the Amisano and Giacomini (2007) t-statistics computed with a serial correlation-robust variance, using a rectangular kernel, $h - 1$ lags (h denotes the forecast horizon), and the small-sample adjustment of Harvey et al. (1997).

* Significance levels of 10%.

** Significance levels of 5%.

*** Significance levels of 1%.

hence, a figure below 1 in the RMSFE ratio, or above 0 in the SCORE difference, shows that a model is outperforming the random walk benchmark in point and density forecasting, respectively.

We will organize the discussion of the results around the following points: (i) comparison with benchmark models (a random walk no-change forecast and an ATSM with stochastic volatility), (ii) the role of the no-arbitrage prior, (iii) the role of time variation in volatility, (iv) the role of dynamics (lags), and (v) the role of the factor structure. The section concludes with a discussion of the stability of our results in different subsamples.

4.5.1 | Comparison with benchmarks

Comparison with random walk

For point forecasts, our results broadly confirm that the RW is a strong forecasting model for US yields, especially so at the long end of the curve. At the short end of the curve (1- and 3-month yields), the models shrinking toward the JSZ-VAR-CSV outperform the RW systematically, with gains up to 17.5% for the 1-month yield and up to 5.5% for the 3-month yield. In some cases, some other specifications based on the JSZ prior (discussed in more detail below) perform even better. Notably, the F-VAR-CSV outperforms the RW also at the 1-year yield. These gains are statistically significant at the shorter forecast horizons. As the maturity of the yields increases, the RW produces better point forecasts, even though the differences are generally small and never significant, ranging between 0% and 8%.

For density forecasts, the JSZ-VAR-CSV produces more accurate forecasts than the RW in most cases. Again, the best results are obtained at the short end of the curve, but positive gains are still present for the medium-term maturities. For the 1- and 3-month yields, the gains in the score are large and significant, ranging from 19% to 45% (note that the difference in log score can be interpreted as a percentage gain in the score). At the 1-year maturity, the gains range between 10% and 20%. At the 2- and 3-year maturities, the gains are typically positive but insignificant. At the long end of the curve, the relative scores are negative but small and insignificant, signaling that the forecasting performance is virtually the same as that of the RW.

Comparison with the DNS-SV ATSM

We assess to what extent the JSZ-VAR-CSV model improves forecast accuracy relative to the dynamic Nelson–Siegel specification of Diebold and Li (2006) with the addition of time-varying volatility in the factors. This model is labeled DNS-SV.

For both point and density forecasts, the DNS-SV model does not fare well compared with the RW benchmark at any horizon. In all cases, the JSZ-VAR-CSV achieves much better forecast accuracy. This finding confirms the difficulty that ATSMs have in outperforming the random walk in forecasting the yield curve originally documented in Duffee (2002) . An inspection of the density forecasts over time reveals that the DNS-SV tends to revert to the mean too quickly compared with the other models, which especially hampers its forecasting performance in the final part of the sample that is characterized by very low yields.[26]

4.5.2 | The role of no-arbitrage priors

Arguably, shrinkage can help in forecasting regardless of the direction in which it is applied, simply because it reduces the problem of overparameterization. It is natural to ask whether the forecasting gains of the JSZ-VAR-CSV model are coming because the coefficients are shrunk toward the JSZ representation or simply because shrinking per se reduces noise in estimation. To check this, one can compare the forecasting performance of the JSZ-VAR-CSV with that of the BVAR-CSV, that is, the VAR with a Minnesota-style prior and CSV. These two models have the same form (they are both VARs with stochastic volatility) and dynamics (they have the same number of lags). Hence, they also have the same likelihood function. The only difference between these two models is in the priors, that is, in the direction in which shrinkage takes place.

For point forecasts, the JSZ-VAR-CSV produces the best forecasts at the 1-month maturity, at all forecast horizons. For the remaining maturities, the BVAR-CSV performs better at short forecast horizons (with gains vs. the JSZ-VAR-CSV up to 4%) whereas the JSZ-VAR-CSV performs better at long forecast horizons (with gains versus the BVAR-CSV up to 11%). Overall, the differences are small.

Turning to density forecasts, the evidence in favor of using the JSZ prior as opposed to the Minnesota prior becomes starker. The JSZ-VAR-CSV outperforms the BVAR-CSV at most horizons and maturities. The only cases in which the

[26]In unreported results based on the same sample of the original Diebold and Li (2006) paper, the model performs much better.

Minnesota prior performs better are the one-step-ahead forecasts for yields with maturity of 2 years or more, and the six- and 12-step-ahead forecasts of the 10-year yields. In these few cases, the largest gain of the BVAR-CSV versus the JSZ-VAR-CSV is of only 2.4% (one-step-ahead forecast of the 1-year yield). In all of the remaining cases, the JSZ-VAR-CSV performs better than the BVAR-CSV, especially at the long end of the curve and at longer forecast horizons, with gains up to 32.5% (12-step-ahead forecast of the 3-month yield). In general, the gains are smaller as we move toward the long end of the curve, and one should also note that at the long end of the curve, the RW still outperforms both of these models, albeit by a very small, insignificant margin.

In summary, these results provide evidence that the forecasting gains do not merely come from the use of shrinkage. The *direction* of shrinkage is just as important.

4.5.3 | The role of stochastic volatility

We now turn to analyzing the role of the assumption of time variation in volatility, by comparing the forecasts from the homoskedastic version (JSZ-VAR) with the heteroskedastic version (JSZ-VAR-CSV) of our model.

It is evident that the addition of stochastic volatility greatly improves the density forecasts, with scores from the heteroskedastic model being in all instances well above those of the homoskedastic version. This pattern changes at the long forecast horizons, but note that this result is driven by a few observations at the beginning of the 2007 financial crisis: at the beginning of this period, the JSZ-VAR-CSV features an extremely low estimated volatility, because most data in this period belong to the Great Moderation and it takes time for the model to learn that the regime has changed (this effect is much more pronounced if one considers 1-year-ahead forecasts). Instead, around this period, the JSZ-VAR fares better because the very fact that it is homoskedastic means it "remembers" the periods of high volatility in the 1980s and 1990s, and this in turn implies wider credible bounds around its point forecasts. The JSZ-VAR produces generally better point forecasts than the JSZ-VAR-CSV, but formal tests of equal accuracy (not reported) reveal that such differences are never statistically significant.

4.5.4 | The role of dynamics

As shown in Equations (1) and (2), a specification assumption in the JSZ model (and any other typical ATSM) is for the factors—and consequently the yields—to be a Markov process. We tried to assess how big of a role this constraint plays. The model JSZ-VAR-CSV (one lag) in the tables is a version of the JSZ-VAR-CSV model in which the VAR lag length is set to 1, rather than the optimally selected length of 3. For point forecasts, the JSZ-VAR-CSV (one lag) model produces virtually identical forecasts compared with the baseline model with three lags.

For density forecasts, results are more mixed. At the short end of the curve, the model with one lag slightly underperforms the baseline model with three lags. For example, using the three-lag model to forecast the 1-month yield at the 12-step-ahead horizon provides a gain of 7.4 %. At the medium and long end of the curve, the forecasts are virtually identical.

This indicates that the Markov assumption may be good for the yields at the long end of the curve, less so for those at the short end, and that the gains we documented at the short end of the curve come at least in part from the fact that the VAR has richer dynamics than the typical one-lag dynamics assumed in term structure models.

4.5.5 | The role of the factor structure

We discussed how shrinking the coefficients toward the JSZ reference model provides forecasting gains. However, note that the JSZ model entails both the assumption of a factor model for the yields, given in Equations (1) and (2) and a set of restrictions on the intercepts and factor loadings of the model, that is, the Riccati equations (3 and 4). Here, we want to evaluate the relative importance of these alternative restrictions on forecast accuracy. In order to do so, we have considered a VAR with a prior that only implements the factor structure assumption, without imposing the restrictions on the loadings (the F-VAR-CSV model).

Looking at the corresponding panels in Tables 2 and 3, it is clear that this model actually has a very good performance, often obtaining the best overall results compared with other variants of the model. For example, the F-VAR-CSV is the only model consistently outperforming the RW at the 1-year maturity in point forecasting. It performs generally better (albeit not by much) than the JSZ-VAR-CSV in point forecasting. In density forecasting, it is overall the best model at the short end of the curve.

This clearly indicates that the cross-equation no-arbitrage restrictions on the loadings only play a minor role in improving forecast accuracy, while most gains come from the imposition of a factor structure. These results are in line with the

argument of Duffee (2011a), who makes the case that because the loadings of the model can be estimated with extremely high precision even if no-arbitrage restrictions are not imposed, the Gaussian no-arbitrage model, absent additional restrictions on risk premia, offers no advantages over a simple unrestricted factor model.

4.5.6 | Subsample analysis

In order to assess the stability of our results throughout the sample, we have computed the loss functions (RMSFE and SCORE) recursively. Results of this exercise are displayed in Figures 4 and 5 for point and density forecasts, respectively. In order to avoid cluttering the graphs, we focus only on three maturities (1 month, 3 month, and 10 years) and two forecast horizons (one step and 12 steps ahead). Unreported results for the remaining combinations show patterns that are in between the ones displayed in these figures.

In Figure 4, the relative RMSFE against the RW is reported. A value below 1 signals that the JSZ-VAR-CSV is outperforming the RW benchmark. Also, as the series depicted is a recursive mean, whenever the series is trending downward, the forecasting performance of the JSZ-VAR-CSV is improving (relative to the RW), whereas when it is trending upward, the forecasting performance is deteriorating. From an inspection of the picture, several conclusions can be drawn.

At the one-step-ahead forecast horizon, the JSZ-VAR-CSV is outperforming the RW throughout the sample, but the relative gains are not stable. Consider, for example, the 3-month yield. The JSZ-VAR-CSV performs well at the beginning of the sample, then the forecasting performance deteriorates starting from the end of the 1990s until the first few years of the 2000s, with the RMSFE ratio versus the RW reaching a value just below 1 around 2002. Then, there is a sharp improvement, after which the RMSFE ratio versus the RW remains steady at around 0.6 for several years, until the 2007 financial crisis ensues, bringing the ratio up again, just below 1. A broadly similar pattern can be found for the 1-month rate, even though the ratios do not deteriorate as much during the first part of the sample. The point forecasts of the 10-year yield show a slow deterioration in forecasting performance, without major jumps. At the long forecast horizon, the relative gains are stable. The JSZ-VAR-CSV underperforms the RW at the beginning of the sample, but then, it starts improving around 1999, and the improvements continue steadily and the ratios eventually stabilize.

Turning to density forecasts, results are displayed in Figure 5. In this figure, we report the difference in the average SCORE between the JSZ-VAR-CSV and the RW. A value above 0 signals that the JSZ-VAR-CSV is outperforming the RW. Also, as the series depicted is a recursive mean, whenever the series is trending upward, the forecasting performance of the JSZ-VAR-CSV is improving (relative to the RW), whereas when it is trending downward, the forecasting performance

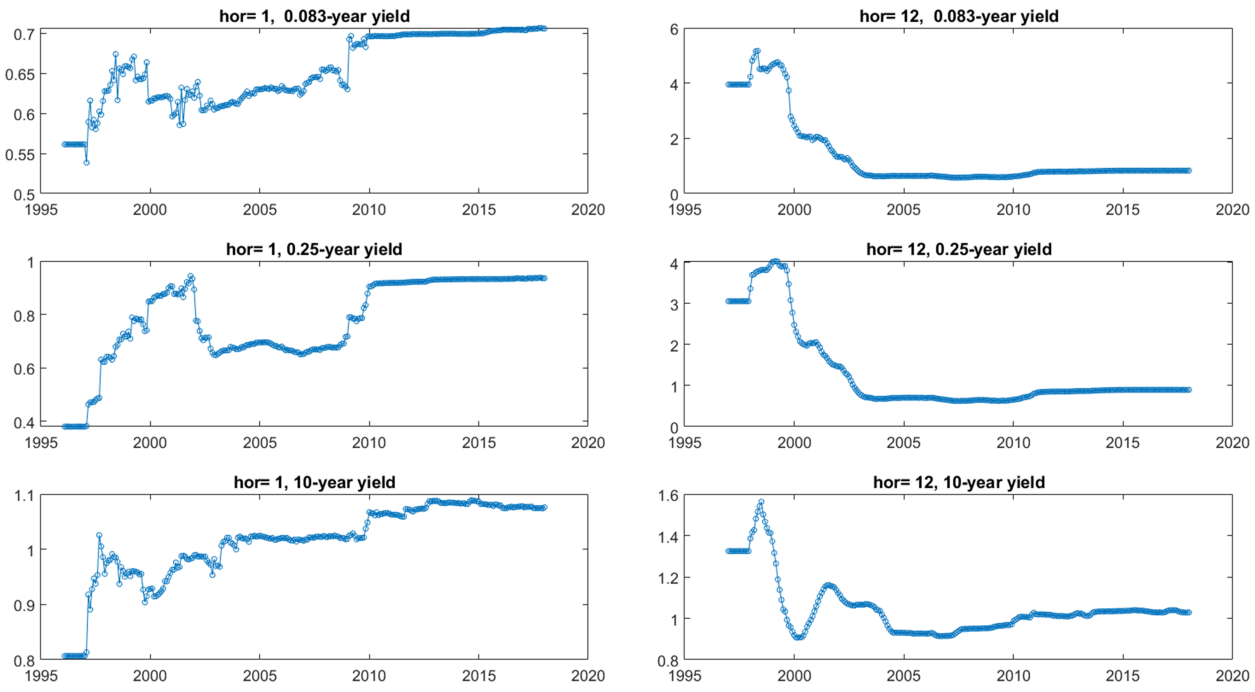

FIGURE 4 Recursive relative RMSFE of the JSZ-VAR-CSV versus the random walk. Recursive means are computed starting from January 1996 [Colour figure can be viewed at wileyonlinelibrary.com]

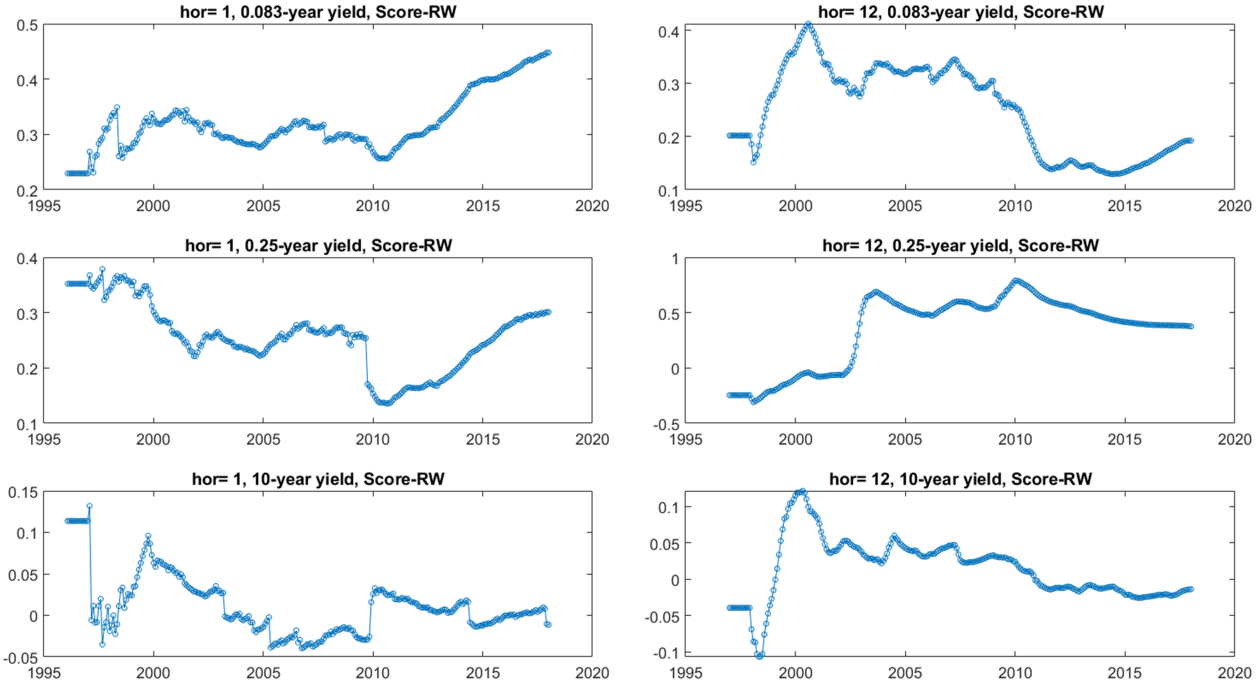

FIGURE 5 Recursive difference in SCORE of the JSZ-VAR-CSV versus the random walk. Recursive differences are computed starting from January 1996. [Colour figure can be viewed at wileyonlinelibrary.com]

is deteriorating. In general, relative SCOREs show more instability than relative RMSFEs; however, it is important to note that they are almost always positive, signaling that the JSZ-VAR-CSV produced the best density forecasts *throughout the sample*. Interestingly, the one-step-ahead forecasting performance generally improves after the financial crisis, whereas the 12-step-ahead performance worsens.

5 | CONCLUSIONS

In this paper, we proposed a way to impose a no-arbitrage ATSM as a prior on a VAR while also allowing for time variation in the error volatilities. As the volatilities of yields move closely together, we imposed a factor structure in which the volatility of each yield is related to a CSV factor, as in Carriero et al. (2016). To shrink the VAR coefficients toward the values implied by an underlying ATSM, we adapted the methodology put forward by Del Negro and Schorfheide (2004). The model toward which VAR coefficients are shrunk is the canonical no-arbitrage model of Joslin et al. (2011).

We provided the conditional posterior distribution kernels of the coefficients and states of the model and developed an MCMC algorithm to draw from their joint posterior. While we applied the proposed model to term structure forecasting, the same approach can be applied to a wide range of alternative models, including DSGE models, and can be considered an extension of the method of Del Negro and Schorfheide (2004) to VARs featuring drifting volatilities with a common factor structure.

By estimating the model using US data on government bond yields covering the period from January 1985 to December 2018, we provided evidence that this method produces competitive forecasts. Compared with a fully fledged ATSM with time-varying volatility, the proposed model consistently produced better point and density forecasts. Compared with a random walk, which is typically a very strong benchmark in forecasting the yields, the model fared consistently better at the short and medium end of the curve and equally well at the long end of the curve.

We have further investigated which of the assumptions that we relaxed are most important to the out-of-sample forecast accuracy of ATSMs. Our findings show that the forecasting gains mainly stem from the partial relaxation of some restrictions that are typical of ATSMs. In particular, we found that the VAR representation might work better because it relaxes the requirement that yields obey a strict factor structure and that the factors follow a Markov process. Instead, we found that the cross-equation no-arbitrage restrictions on the loadings only have a marginal role, in line with Duffee (2011a).

ACKNOWLEDGMENTS

We thank the editor, Michael McCracken, three anonymous referees, Caio Almeida, Carlo Altavilla, Gianni Amisano, Domenico Giannone, Joe Haubrich, Giorgio Primiceri, Minchul Shin, Rafael Wouters, Jonathan Wright, participants at the IAAE Annual Conference, the ECB forecasting workshop, the NBER-NSF Seminar on Bayesian Inference in Econometrics and Statistics, and seminar participants at the ECB and ECARES for useful comments and suggestions. Jared Steinberg and Tommaso Tornese provided excellent research assistance. The views expressed herein are solely those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of Cleveland or the Federal Reserve System.

OPEN RESEARCH BADGES

This article has been awarded Open Data Badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. Data is available at [http://qed.econ.queensu.ca/jae/forthcoming/carriero-clark-marcellino/].

REFERENCES

Adolfson, M., Linde, J., & Villani, M. (2007). Forecasting performance of an open economy DSGE model. *Econometric Reviews*, *26*, 289–328.

Almeida, C., & Vicente, J. (2008). The role of no-arbitrage on forecasting: Lessons from a parametric term structure model. *Journal of Banking and Finance*, *32*, 2695–2705.

Amisano, G., & Giacomini, R. (2007). Comparing density forecasts via weighted likelihood ratio tests. *Journal of Business and Economic Statistics*, *25*, 177–190.

Ang, A., Boivin, J., Dong, S., & Loo-Kung, R. (2011). Monetary policy shifts and the term structure. *Review of Economic Studies*, *78*, 429–457.

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

Banbura, M., Giannone, D., & Reichlin, L. (2010). Large Bayesian vector autoregressions. *Journal of Applied Econometrics*, *25*, 71–92.

Bauer, M. D. (2018). Restrictions on risk prices in dynamic term structure models. *Journal of Business and Economic Statistics*, *36*, 196–211.

Bernanke, B. (2004). The great moderation. In *Remarks at the Meetings of the Eastern Economic Association*, Washington, DC.

Carriero, A. (2011). Forecasting the yield curve using priors from no-arbitrage affine term structure models. *International Economic Review*, *52*, 425–459.

Carriero, A., Clark, T. E., & Marcellino, M. (2015). Bayesian VARs: Specification choices and forecast accuracy. *Journal of Applied Econometrics*, *30*, 46–73.

Carriero, A., Clark, T. E., & Marcellino, M. (2016). Common drifting volatility in large Bayesian VARs. *Journal of Business and Economic Statistics*, *34*, 375–390.

Carriero, A., Kapetanios, G., & Marcellino, M. (2012). Forecasting government bond yields with large Bayesian VARs. *Journal of Banking and Finance*, *36*, 2026–2047.

Chib, S., & Ergashev, B. (2009). Analysis of multifactor affine yield curve models. *Journal of the American Statistical Association*, *104*, 1324–1337.

Chib, S., & Kang, K. H. (2014). Change-points in affine arbitrage-free term structure models. *Journal of Financial Econometrics*, *11*, 302–334.

Christensen, J. H. E., Diebold, F. X., & Rudebusch, G. D. (2011). The affine arbitrage-free class of Nelson-Siegel term structure models. *Journal of Econometrics*, *164*, 4–20.

Clark, T., & McCracken, M. W. (2015). Nested forecast model comparisons: A new approach to testing equal accuracy. *Journal of Econometrics*, *186*, 160–177.

Clark, T. E., & McCracken, M. W. (2011). Testing for unconditional predictive ability. In M. P. Clements, & D. F. Hendry (Eds.), *Oxford handbook of economic forecasting*: Oxford University Press.

Cogley, T., & Sargent, T. (2005). Drifts and volatilities: Monetary policies and outcomes in the post-WWII US. *Review of Economic Dynamics*, *8*, 262–302.

Creal, D. D., & Wu, J. C. (2015). Estimation of term structure models with spanned or unspanned stochastic volatility. *Journal of Econometrics*, *185*, 60–81.

Creal, D. D., & Wu, J. C. (2017). Monetary policy uncertainty and economic fluctuations. *International Economic Review*, *58*, 1317–1354.

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

Del Negro, M., & Schorfheide, F. (2004). Priors from general equilibrium models for VARs. *International Economic Review*, *45*, 643–673.

Diebold, F., & Mariano, R. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics*, *13*, 253–63.

Diebold, F. X., & Li, C. (2006). Forecasting the term structure of government bond yields. *Journal of Econometrics*, *130*, 337–364.

Doan, T., Litterman, R., & Sims, C. (1984). Forecasting and conditional projection using realistic prior distributions. *Econometric Reviews*, *3*, 1–100.

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

Duffee, G. (2011a). Forecasting with the term structure: The role of no-arbitrage restrictions. Working paper, Johns Hopkins University.

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

Duffee, G., & Stanton, R. (2012). Estimation of dynamic term structure models. *Quarterly Journal of Finance*, *2*, 1–51.
Duffie, D., & Kan, R. (1996). A yield-factor model of interest rates. *Mathematical Finance*, *6*, 379–406.
Geweke, J., & Amisano, G (2010). Comparing and evaluating Bayesian predictive distributions of asset returns. *International Journal of Forecasting*, *26*, 16–230.
Giacomini, R., & Ragusa, G. (2014). Theory-coherent forecasting. *Journal of Econometrics*, *182*, 145–155.
Giannone, D., Lenza, M., & Primiceri, G. (2015). Prior selection for vector autoregressions. *Review of Economics and Statistics*, *97*, 436–451.
Gürkaynak, R. S., Sack, B., & Wright, J. H. (2007). The U.S. Treasury yield curve: 1961 to the present. *Journal of Monetary Economics*, *54*, 2291–2304.
Hamilton, J., & Wu, J. C. (2012). Identification and estimation of Gaussian affine term structure models. *Journal of Econometrics*, *168*, 315–331.
Hamilton, J., & Wu, J. C. (2014). Testable implications of affine term structure models. *Journal of Econometrics*, *178*, 231–242.
Harvey, D., Leybourne, S., & Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of Forecasting*, *13*, 281–291.
Hautsch, N., & Ou, Y. (2012). Analyzing interest rate risk: Stochastic volatility in the term structure of government bond yields. *Journal of Banking and Finance*, *36*, 2988–3007.
Hong, Y., Li, H., & Zhao, F. (2004). Out-of-sample performance of discrete-time spot interest rate models. *Journal of Business and Economic Statistics*, *22*, 457–473.
Johannsen, B. K., & Mertens, E. (2021). A time series model of interest rates with the effective lower bound. *Journal of Money, Credit, and Banking*, forthcoming.
Joslin, S., Le, A., & Singleton, K. J. (2013a). Why Gaussian macro-finance term structure models are (nearly) unconstrained factor-VARs. *Journal of Financial Economics*, *109*, 604–622.
Joslin, S., Le, A., & Singleton, K. J. (2013b). Gaussian macro-finance term structure models with lags. *Journal of Financial Econometrics*, *11*, 581–609.
Joslin, S., Priebsch, M., & Singleton, K. J. (2014). Risk premiums in dynamic term structure models with unspanned macro risks. *Journal of Finance*, *69*, 1197–1233.
Joslin, S., Singleton, K. J., & Zhu, H. (2011). A new perspective on Gaussian dynamic term structure models. *Review of Financial Studies*, *24*, 926–970.
Kadiyala, K. R., & Karlsson, S. (1997). Numerical methods for estimation and inference in Bayesian VAR-models. *Journal of Applied Econometrics*, *12*, 99–132.
Koop, G. M. (2013). Forecasting with medium and large Bayesian VARs. *Journal of Applied Econometrics*, *28*, 177–203.
Krippner, L. (2013). Measuring the stance of monetary policy in zero lower bound environments. *Economics Letters*, *118*, 135–138.
Litterman, R. (1986). Forecasting with Bayesian vector autoregressions—Five years of experience. *Journal of Business and Economic Statistics*, *4*, 25–38.
Nelson, C. R., & Siegel, A. F. (1987). Parsimonious modeling of yield curves. *Journal of Business*, *60*, 473–489.
Robertson, J. C., & Tallman, E. W. (1999). Vector autoregressions: Forecasting and reality. *Economic Review-Federal Reserve Bank of Atlanta*, *84*(Q1), 4–18.
Robertson, J. C., Tallman, E. W., & Whiteman, C. H. (2005). Forecasting using relative entropy. *Journal of Money, Credit and Banking*, *37*, 383–401.
Shin, M., & Zhong, M. (2017). Does realized volatility help bond yield density prediction? *International Journal of Forecasting*, *33*, 373–389.
Sims, C. (1993). A nine-variable probabilistic macroeconomic forecasting model. In J. H. Stock, & M. W. Watson (Eds.), *Business cycles, indicators and forecasting*: University of Chicago Press, pp. 179–204.
Sims, C., & Zha, T. (1998). Bayesian methods for dynamic multivariate models. *International Economic Review*, *39*, 949–68.
Vasicek, O. (1977). An equilibrium characterization of the term structure. *Journal of Financial Economics*, *5*, 177–188.
Waggoner, D. F., & Zha, T. (1999). Conditional forecasts in dynamic multivariate models. *The Review of Economics and Statistics*, *81*, 639–651.
Wu, J. C., & Xia, F. D. (2016). Measuring the macroeconomic impact of monetary policy at the zero lower bound. *Journal of Money, Credit, and Banking*, *48*, 253–291.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Carriero A, Clark TE, Marcellino M (2021). No-arbitrage priors, drifting volatilities, and the term structure of interest rates. *Journal of Applied Econometrics*. 2021;36:495–516. https://doi.org/10.1002/jae.2828