DISCUSSION PAPER SERIES

No. 9848

NO ARBITRAGE PRIORS, DRIFTING VOLATILITIES, AND THE TERM STRUCTURE OF INTEREST RATES

Andrea Carriero, Todd Clark and
Massimiliano Marcellino

INTERNATIONAL MACROECONOMICS

Centre for Economic Policy Research

www.cepr.org

Available online at: **www.cepr.org/pubs/dps/DP9848.php**

ISSN 0265-8003

NO ARBITRAGE PRIORS, DRIFTING VOLATILITIES, AND THE TERM STRUCTURE OF INTEREST RATES

Andrea Carriero, Queen Mary, University of London
Todd Clark, Federal Reserve Bank of Cleveland
Massimiliano Marcellino, IGIER, Università Bocconi and CEPR

Discussion Paper No. 9848
March 2014

Centre for Economic Policy Research
77 Bastwick Street, London EC1V 3PZ, UK
Tel: (44 20) 7183 8801, Fax: (44 20) 7183 8820
Email: cepr@cepr.org, Website: www.cepr.org

CEPR Discussion Paper No. 9848

March 2014

ABSTRACT

No Arbitrage Priors, Drifting Volatilities, and the Term Structure of Interest Rates*

We propose a method to produce density forecasts of the term structure of government bond yields that accounts for (i) the possible mispecification of an underlying Gaussian Affine Term Structure Model (GATSM) and (ii) the time varying volatility of interest rates. For this, we derive a Bayesian prior from a GATSM and use it to estimate the coefficients of a BVAR for the term structure, specifying a common, multiplicative, time varying volatility for the VAR disturbances. Results based on U.S. data show that this method significantly improves the precision of point and density forecasts of the term structure. While this paper focuses on term structure modelling, the proposed method can be applied for a wide range of alternative models, including DSGE models, and is a generalization of the method of Del Negro and Schorfheide (2004) to VARs featuring drifting volatilities. The method also generalizes the model of Giannone et al. (2012), by specifying hierarchically not only the prior variance but also the prior mean of the VAR coefficients. Our results show that both time variation in volatilities, and a hierarchical specification for the prior means, improve model fit and forecasting performance.

JEL Classification: C32, C53 and G17
Keywords: density forecasting, no arbitrage, stochastic volatility and term structure

Andrea Carriero
Queen Mary, University of London
Mile End Road
London E1 4NS

Email: a.carriero@qmul.ac.uk

For further Discussion Papers by this author see:
www.cepr.org/pubs/new-dps/dplist.asp?authorid=160379

Todd Clark
Federal Reserve Bank of Cleveland
P.O. Box 6387
Cleveland, OH 44101-1387
USA

Email: todd.clark@clev.frb.org

For further Discussion Papers by this author see:
www.cepr.org/pubs/new-dps/dplist.asp?authorid=170994

Massimiliano Marcellino
Department of Economics
Bocconi University
Via Roentgen 1
20136, Milano
ITALY

Email:
massimiliano.marcellino@unibocconi.it

*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. We thank Caio Almeida, Carlo Altavilla, Gianni Amisano, Domenico Giannone, Joe Haubrich, Giorgio Primiceri, Minchul Shin, Jonathan Wright, and seminar participants at the ECB and ECARES for useful comments and suggestions. The usual disclaimers apply.

Submitted 22 January 2014

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, e.g. Duffee (2002), Ang and Piazzesi (2003), Diebold and Li (2006), Almeida and Vicente (2008), Christensen, Diebold and Rudebusch (2011), Carriero (2011), and Carriero, Kapetanios and Marcellino (2012). All these contributions have focused on point forecasts of the yield curve, but assessing the whole predictive distribution of the yield curve is more important for the success of portfolio and risk management strategies. Hong, Li and Zhao (2004) make a relevant contribution in this context, finding that modelling changes in the variance is important for interest rate density forecasting. Shin and Zhong (2013) find a similar result using realized volatility.

In this paper we focus on forecasting the whole density distribution of future term structures. As the time series of interest rates typically feature comovement and heteroskedasticity, having a joint dynamic model featuring time variation in volatility is key in order to produce reliable density forecasts. Therefore, we propose the use of a Bayesian Vector Autoregressive model whose disturbances follow a stochastic volatility process driven by a single multiplicative factor. Such a model has the advantage that, conditional on the latent volatility factor, it admits a Normal-Inverse Wishart naturally conjugate prior. In addition, we use a particular prior for the model parameters, based on a no arbitrage model.

Joslin, Singleton, and Zhu (2011) have provided a representation of Gaussian Affine Term Structure Models (GATSM) which makes clear that, in the absence of additional restrictions on risk premia, no-arbitrage restrictions per se do not affect the dynamics of the factors governing the yield curve, and therefore do not affect the forecasts of such factors. However, they also show that no-arbitrage restrictions do affect the estimation of the loadings (and therefore affect the mapping from the factor forecasts to the yields forecasts), as well as the variances of the errors in both the measurement and transition equations of the model (and therefore the density forecasts), and conclude that "the role of no-arbitrage restrictions is an empirical issue."[1]

[1] Duffee (2011) argues that since 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 regression-based approach. This argument does not apply if one considers comparing forecasts from a GATSM against forecasts from an unrestricted Vector Autoregression (VAR), as is the case in this paper. Indeed, beyond the no arbitrage restrictions, the factor structure inherent in a GATSM restricts the data more than an unrestricted VAR and therefore can provide gains in forecast accuracy. Finally, note that while we do not pursue this route in this paper, our

Several papers have analyzed this issue, with mixed results. Duffee (2002) and Ang and Piazzesi (2003) have shown that beating a random walk with a traditional no arbitrage GATSM is difficult. Evidence in favour of using no arbitrage restrictions is provided by Almeida and Vicente (2008), Christensen, Diebold and Rudebusch (2011), and Carriero and Giacomini (2011). Interestingly, in the case of Almeida and Vicente (2008), one of the reasons for the difference in results with respect to the rest of the literature is that they consider models with stochastic volatility, while most of the literature only adopts Gaussian models in forecasting applications.

One of the reasons behind the mixed results concerning the usefulness of no arbitrage restrictions might be related to the fact that the assumption of absence of arbitrage - which is per se reasonable in well developed markets - needs nonetheless to be translated into a set of restrictions to impose on a particular model. This process requires a set of additional specification assumptions, which are not necessarily holding in the data. For this reason, we propose using a no arbitrage model as a prior rather than as a set of sharp restrictions on the model parameters, which allows us to take into account the potential misspecification of the model. The use of a no-arbitrage model as a prior will result in shrinkage of the posterior distributions of the parameters of a Vector Autoregression — and consequently of its density forecasts — in an economically-meaningful direction, which might (and in our application does) improve the forecasting performance with respect to both a fully restricted no-arbitrage model, and a fully unrestricted VAR.[2]

Carriero (2011) conducts a similar exercise, but only under the hypotheses of Gaussianity and conditional homoskedasticity of the yields, mild assumptions for point forecasting but likely inadequate for density forecasting. In particular, using the methodology put forward by Del Negro and Schorfheide (2004), Carriero (2011) develops a model for the term structure in which the no arbitrage restrictions are imposed as prior information rather than dogmatically, and shows that once the misspecification of the model is properly taken into account the point forecasting performance can improve substantially.

In this paper we extend this methodology to the case of a VAR featuring stochastic

framework lends itself naturally to imposing additional restrictions on the dynamics of risk premia, which in turn would impact the dynamics of the factors. As shown by Duffee (2011) such restrictions can further improve the forecast accuracy of a GATSM.

[2] Our approach is in spirit similar to the relative entropy procedures of Robertson, Tallman, and Whiteman (2005), and Giacomini and Ragusa (2011). In the entropy approach, the forecasts are "tilted" towards 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. In our 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.

volatility. As the volatilities of a panel of yields move closely together, we impose on them a factor structure where the volatility of each yield is ultimately related to a common stochastic volatility factor, as in Carriero, Clark and Marcellino (2012). This approach shrinks the point and density forecasts towards values consistent with both the validity of the GATSM, and time variation in the volatilities.

As we shall see, such a modelling choice results in a clear improvement in density forecasting performance. In particular, the proposed model produces better density forecasts than a model in which the estimates are shrunk towards a GATSM but the time variation in volatility is shut down. It also produces better point and density forecasts than the case in which the GATSM is imposed exactly on the data, and than a random walk, which is regarded as a very competitive benchmark in yield curve forecasting.[3] Further analysis reveals that the most relevant feature of the GATSM prior to improve forecast accuracy is the imposition of a factor structure on the yields, while the additional no-arbitrage restrictions imposed on the loadings provide only marginal improvements -if any- in forecasting, a result broadly in line with Duffee (2011a).

Moreover, while Carriero's (2011) prior specification is based on the model by Ang and Piazzesi (2003), which is the discrete time version of the specification proposed by Duffie and Kan (1996), here we consider the new canonical form of no arbitrage models introduced by Joslin, Singleton, and Zhu (2011). This new canonical representation presents very important advantages in the computation of the likelihood, because it allows one to disentangle the role played by the structural parameters in the risk neutral and physical measure, which in turn allows one to factorize the likelihood and concentrate out some parameters, reducing dramatically the difficulties typically arising in the search for the global optimum of the likelihood. In our exercise, the factorization will allow us to concentrate out some of the parameters of the model, reducing dramatically the number of coefficients to estimate and yielding better mixing properties and faster computation time at a relatively small cost.

It is worth stressing that while in this paper we consider term structure forecasting, this is only one of the possible applications of our proposed method. It can be applied for a wide range of alternative models, including DSGE models, and can be considered as an extension of the method of Del Negro and Schorfheide (2004) to VARs featuring drifting volatilities. Our proposed model also nests the one of Giannone, Lenza and Primiceri

[3] To assess the relative merits of the no arbitrage restriction, we also include in the comparison a BVAR with stochastic volatility where the prior means are shrunk towards a random walk, resembling a version of the Minnesota prior. We find that such a model is also systematically outperformed by our proposed model in both density and point forecasting.

(2012), as in their approach the model is homoskedastic and only the prior variance of the VAR coefficients is specified in a hierarchical fashion, while in our approach the model is heteroskedastic, and both the prior variance and the prior mean of the VAR coefficients are specified hierarchically. Our results show that a hierarchical specification for the prior means does help in forecasting, highlighting the fact that not only shrinkage per se, but also the direction towards which shrinkage takes place, can be helpful.

The paper is organized as follows. Section 2 describes the no arbitrage model used in the paper. Section 3 discusses the priors, derives the conditional posteriors (with additional details in the Appendix), and briefly describes the other BVAR models to which we compare the results from our proposed specification. Section 4 discusses the MCMC implementation. Section 5 presents our U.S.-based evidence, including both a full-sample evaluation and an out-of sample forecasting assessment. Section 6 summarizes the main results and concludes. Finally, an Appendix provides additional details.

2 The JSZ Canonical Affine Term Structure Model

Since the seminal work of Vasicek (1977) a large part of research has focused on Gaussian Affine Term Structure Models (GATSM). Prominent contributions in this tradition include Duffie and Kan (1996), Dai and Singleton (2000), Duffee (2002), and Ang and Piazzesi (2003). Traditional GATSM entail a high level of nonlinearity that makes the estimation extremely difficult and often unreliable. For example Duffee (2009) and Duffee and Stanton (2012) show that there are considerable problems in reaching the global optimum, and Hamilton and Wu (2012) show that even some commonly used models are not identified. Some recent literature has successfully addressed this issue. Hamilton and Wu (2012) propose a strategy to estimate such models using a series of transformations and OLS estimation. Christensen, Diebold and Rudebusch (2011) proposed a term structure model based on the Nelson and Siegel (1987) exponential framework, featuring the important extension that no arbitrage is imposed on the cross section of yields.

Joslin, Singleton and Zhu (2011) (JSZ) recently proposed a representation of GATSM equivalent to the canonical representation of Duffie and Kan (1996), but parametrized in such a way that estimation is considerably simplified. Under such representation, a convenient factorization of the likelihood arises, which allows fast convergence of standard ML to the global optimum, whereas estimation of models represented as in Duffie and Kan (1996) is typically problematic. The computational advantage of the JSZ approach stems from the fact that the particular rotation of the model they use makes evident the fact that the

dynamics of the factors driving the cross section of yields is completely independent from the assumption of absence of arbitrage.[4]

The method proposed by JSZ is particularly well suited for our purposes, because the estimation of the JSZ structural parameters will be a step in a large Markov Chain Monte Carlo (MCMC) algorithm, and the factorization they provide will allow us to concentrate out some of the parameters of the model, thereby reducing the number of coefficients to estimate, yielding better mixing properties and faster estimation. In what follows we summarize the representation proposed by JSZ; the interested reader can find additional details on the derivation in Appendix A or in JSZ.

Term structure models assume that the evolution of yields over time is driven by some factors, which can be either observable or latent. Then, given the factor dynamics, the assumption of no arbitrage implies a set of restrictions on the movements of the yields in the cross section. In the canonical Duffie and Kan (1996) representation, the evolution of the n factors (a n-dimensional state vector S_t) is given by:

$$\Delta S_t = K_{0S}^{\mathbb{P}} + K_{1S}^{\mathbb{P}} S_{t-1} + \Sigma_S \varepsilon_t^{\mathbb{P}} \tag{1}$$

$$\Delta S_t = K_{0S}^{\mathbb{Q}} + K_{1S}^{\mathbb{Q}} S_{t-1} + \Sigma_S \varepsilon_t^{\mathbb{Q}} \tag{2}$$

$$r_t = \rho_{0S} + \rho_{1S} S_t \tag{3}$$

where $\mathbb{Q}$ and $\mathbb{P}$ denote the risk neutral and physical measures of probability, r_t is the short term rate, Σ_S is the Cholesky factor of the conditional variance of the states, and the errors are i.i.d. Gaussian random variables with mean 0 and variance 1.

Under the $\mathbb{Q}$ probability measure, prices are a martingale, which resembles an hypothetical situation in which investors are risk neutral. Under the $\mathbb{P}$ 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. Absence of arbitrage and the existence of the equivalent martingale measure $\mathbb{Q}$ are equivalent conditions (Harrison and Kreps, 1979). Conversion from the $\mathbb{P}$ to the $\mathbb{Q}$ measure can be achieved using a variable transformation described by a Radon-Nikodym derivative that, together with the risk free rate, forms the pricing kernel. Absence of arbitrage and the existence of the pricing kernel are also equivalent conditions.[5]

[4] As JSZ stress, because their representation and Duffie and Kan's (1996) representation are equivalent, they both have this feature. However, in the latter representation this is less evident because it is hidden by the rotation used.

[5] 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.

It is important to distinguish the assumption of absence of arbitrage and the additional specification restrictions inherent in a GATSM. In particular, any further assumption beyond the existence of the equivalent martingale measure (or equivalently the existence of a pricing kernel) goes beyond the assumption of no-arbitrage per se. Of course, practical implementation of GATSM requires to specify a distribution for the pricing kernel and a law of motion of the factors, but these are additional assumptions going beyond the mere absence of arbitrage, and as such they can introduce misspecification. For example, the use of a VAR(1) for the law of motion of the factors under the $\mathbb{P}$ measure is an assumption completely unrelated to the absence of arbitrage, and it can introduce misspecification in the model.[6] With regards to the pricing kernel, most papers in this tradition assume a log-normal distribution, an assumption that provides tractability but can induce misspecification.

Of course, one way to solve this problem is to search for a model better explaining the yields dynamics and their cross sectional variation. However, the joint hypothesis problem can not be avoided, and the search for a better specification poses serious problems, as the use of more complex specifications may actually worsen the misspecification of the model.

In this paper, we propose an alternative route, which starts from the acknowledgement that any term structure model will suffer some degree of misspecification. Instead, a Vector Autoregression (VAR) — provided its dynamics is sufficiently rich — is more likely to offer an accurate representation of the data. Therefore we propose to model yields using a VAR, while at the same time shrinking the VAR parameters in the direction of parameters implied by a GATSM. Importantly, in order to avoid misspecification, we do not impose these restrictions sharply, but we allow for some noise around them. This amounts to using the moments implied by the validity of the GATSM on the yields as prior information on the VAR coefficients.

Returning to the specification of the GATSM on which the VAR coefficient prior will be based, the model-implied yields on a set of zero-coupon bonds of maturity $\tau = 1, ..., N$ are

[6] For example, Duffee (2011b) shows that it is entirely possible for the factors to follow richer dynamics in the physical measure than in the risk neutral measure, and that this translates to the presence of hidden factors which -while not useful in explaining the cross-section of yields- can help in explaining their dynamics. Similarly, Joslin, Priebsch, and Singleton (2012) show that a VAR representation (under the physical measure) including measures of real economic activity and inflation captures better the dynamics of the term structure. In this paper we illustrate the proposed approach using the simpler framework offered by yields-only models, but our approach can be naturally extended to models allowing for macroeconomic factors.

an affine function of the state S_t:

$$\tilde{y}_t = A(\Theta_S^{\mathbb{Q}}) + B(\Theta_S^{\mathbb{Q}})S_t, \tag{4}$$

where y_t^* is a $N \times 1$ vector of yields, $A(\Theta_S^Q)$ and $B(\Theta_S^Q)$ are $N \times 1$ and $N \times n$ coefficient matrices which are functions of the deep parameters $\Theta_S^Q = \{K_{0S}^Q, K_{1S}^Q, \Sigma_S, \rho_{0S}, \rho_{1S}\}$ through a set of Riccati equations. Here the use of the symbol $\tilde{}$ highlights that the yields are assumed to be perfectly priced by the model, i.e. they contain no measurement error. To allow for measurement error, define y_t as a $N \times 1$ vector of observable yields:

$$y_t = A(\Theta_S^{\mathbb{Q}}) + B(\Theta_S^{\mathbb{Q}})S_t + \Sigma_y \varepsilon_t^y, \tag{5}$$

where ε_t^y is a vector of *i.i.d.* $N(0,1)$ measurement errors, and Σ_y is a lower triangular matrix. For estimation, equations (1) and (5) form a Gaussian state space model, for which the likelihood is readily available via the Kalman filter.

JSZ derive the following equivalent representation for equations (1) and (5):

$$\Delta P_t = K_{0P}^{\mathbb{P}} + K_{1P}^{\mathbb{P}} P_{t-1} + \Sigma_P \varepsilon_t^{\mathbb{P}} \tag{6}$$

$$y_t = A_p(\Theta_P^{\mathbb{Q}}) + B_p(\Theta_P^{\mathbb{Q}})P_t + \Sigma_y \varepsilon_t^y. \tag{7}$$

In (6) and (7), $P_t = W\tilde{y}_t$ are n linear combinations of the N perfectly priced yields (therefore they are portfolios of yields), and Σ_P is the Cholesky factor of their conditional variance. All the coefficients appearing in A_p, B_p are ultimately a function of the deep coefficients of the model $\Theta_P^{\mathbb{Q}} = \{k_\infty^{\mathbb{Q}}, \lambda^{\mathbb{Q}}, \Sigma_P\}$, where $\lambda^{\mathbb{Q}}$ are the (ordered) eigenvalues of $K_{1S}^{\mathbb{Q}}$ and $k_\infty^{\mathbb{Q}}$ is a scalar related to the long run mean of the short rate under the risk neutral measure. Moreover, note that the parameters in $K_{0P}^{\mathbb{P}}$ and $K_{1P}^{\mathbb{P}}$ enter only in the transition equation. Details of the transformation leading to the system (6) and (7) can be found in Appendix A or in JSZ.

The computational benefits from using the JSZ normalization arise from the observation that the least-squares projection of the observable factors $P_t^o = Wy_t$ onto their lagged values will nearly recover the ML estimates of $K_{0P}^{\mathbb{P}}$ and $K_{1P}^{\mathbb{P}}$ to the extent that $P_t^o \approx P_t$ (and the best approximation is given by choosing W using principal components). As we will use the model as a reference point towards which estimates of a Vector Autoregression (VAR) are shrunk, concentrating out $K_{0P}^{\mathbb{P}}$ and $K_{1P}^{\mathbb{P}}$ by estimating them via OLS in a preliminary step is harmless.[7] Therefore, in a model with 3 factors, there are in total $3+1+6+N = 10+N$

[7] Strictly speaking this concentration is exact only if one assumes that the yields are measured without errors. However, as noted by JSZ, the choice of principal components weights ensures that $P_t^o \approx P_t$, and this

parameters to be estimated: the 3 eigenvalues in $\lambda^{\mathbb{Q}}$, $k_{\infty}^{\mathbb{Q}}$, the 6 elements of Σ_P, and the N elements on the diagonal of Σ_y.[8] We collect these in the vector:

$$\theta = (\lambda^{\mathbb{Q}}, k_{\infty}^{\mathbb{Q}}, \Sigma_P, \Sigma_y). \tag{8}$$

As an additional advantage, the OLS estimates of the error variances $\hat{\Sigma}_P$ and $\hat{\Sigma}_y$ based on the observed yields y_t and portfolios P_t^o provide a natural initial guess for Σ_P and Σ_y.

For future reference it is convenient to compute the moments of the yields under the state space model in (6)-(7), for a given θ. The moments of y_t implied by the state space system are derived in Appendix B. They are:

$$E[y_t y_t'] = (A_p + B_p\bar{P})(A_p + B_p\bar{P})' + B_p\Sigma_f B_p' + \Sigma_y\Sigma_y', \tag{9}$$

and:

$$E[y_t y_{t-h}'] = (A_p + B_p\bar{P})(A_p + B_p\bar{P})' + B_p(K_{1P}^P + I)^h\Sigma_f B_p', \tag{10}$$

where $\bar{P} = E[P_t] = -{K_{1P}^{\mathbb{P}}}^{-1}K_{0P}^{\mathbb{P}}$ and $\Sigma_f = E[f_t f_t']$ solves the Lyapunov equation $\Sigma_f = (K_{1P}^{\mathbb{P}} + I)\Sigma_f(K_{1P}^{\mathbb{P}} + I)' + \Sigma_P\Sigma_P'$.

In our methodology, the moments (9) and (10) will be used to form a prior for a Vector Autoregression. It is important to stress that our prior is given by the whole GATSM, and not only by the set of no-arbitrage restrictions. The GATSM imposes: i) a factor structure for the yields, ii) a particular dynamic structure for the factors under the $\mathbb{P}$ measure (a VAR(1) homoskedastic process), and iii) a set of restrictions on the observation equation of the system. Therefore in our setup the imposition of the GATSM as a prior can produce gains in forecast accuracy, even though, as noted by Joslin, Singleton, and Zhu (2011), the no-arbitrage restrictions per se do not affect the dynamics of the factors governing the yield curve, and even though, as argued by Duffee (2011a), imposing restrictions on the observation equation do not provide relevant gains in efficiency. Indeed a Vector Autoregression is a more general representation of the data than a GATSM, and therefore the imposition of the GATSM as a prior rather than as a set of sharp parameter restrictions can provide gains in forecast accuracy. Imposing restrictions may create enough model misspecification to overwhelm any gains from parsimony and harm forecast accuracy. Instead using Bayesian shrinkage to push the model toward the restrictions but not impose them on the

in turn ensures that the concentration is nearly exact. Indeed, Figure 1 in JSZ shows that the model-implied filtered P_t are nearly identical to their observable counterpart P_t^o. As we use this model as a prior, the cost of such (slightly) inexact concentration is largely offset by the benefit of achieving much better mixing properties of the algorithm.

[8] If one were not concentrating out the parameters in $K_{0P}^{\mathbb{P}}$ and $K_{1P}^{\mathbb{P}}$ this would imply having to estimate 12 more parameters.

data may be less likely to create enough misspecification to overwhelm gains from shrinkage (equivalently, parsimony).

Finally, it is worth stressing that the prior derived from (9) and (10) will of course be Gaussian and homoskedastic. Such a prior will be imposed on a Vector Autoregression featuring stochastic volatility, thereby shrinking the posterior distributions of the coefficients towards a homoskedastic no-arbitrage model. While the shrinkage of the point estimates towards the values implied by the GATSM is obviously desirable, the shrinkage towards a homoskedastic representation can be thought as somewhat less appealing. However, one needs to bear in mind that the posterior of the VAR will allow for stochastic volatility, and therefore in practice estimates can (and do) produce drifting volatilities. Alternatively, one could think of using a prior which is already based on a model featuring drifting volatilities, so that there would not be shrinkage towards a homoskedastic representation. To that end, there are broad classes of no arbitrage models with stochastic volatility available, which could be potentially used as a prior. However, because of the absence of Gaussianity, the moments (9) and (10) would no longer be sufficient statistics, and none of these models could produce a manageable conjugate prior for a VAR. The implementation of a heteroskedastic prior would be possible in principle, but it is in practice unmanageable.[9]

3 Vector Autoregression with no arbitrage prior and common stochastic volatility

The second ingredient of our methodology is the specification of a VAR with drifting volatilities for the N yields on government bonds. We use the specification proposed by Carriero, Clark and Marcellino (2012), which they apply to VARs for macroeconomic variables:

$$y_t = \Phi_0 + \Phi_1 y_{t-1} + ... + \Phi_p y_{t-p} + u_t. \quad (11)$$

$$u_t = \lambda_t^{0.5} \epsilon_t, \ \epsilon_t \sim N(0, V), \quad (12)$$

$$\log(\lambda_t) = \phi_0 + \phi_1 \log(\lambda_{t-1}) + \nu_t, \ \nu_t \sim \text{iid } N(0, \phi_2). \quad (13)$$

In our context, y_t is a vector of yields of different maturities, and there is a single volatility process λ_t that is common to all yields, and drives the time variation in the entire variance-covariance matrix of the VAR errors. In order to achieve identification, we set the initial condition of the CSV process to $\lambda_1 = 1$. We group the parameters governing the dynamics of λ_t in the vector $\phi = (\phi_0, \phi_1, \phi_2)$. The scaling matrix V allows the variances of

[9] In particular, the implementation of such a prior would require repeated simulation of artificial data sets from the no arbitrage model used as prior.

each variable to differ by a factor that is constant over time. The variance of u_t is defined as $\Sigma_t = \lambda_t V$.

The assumption of common stochastic volatility 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 data set we use in our empirical application (seven zero-coupon unsmoothed yields) there is a strong commonality, with the first principal component explaining 89% of the individual volatilities of the yields.[10] Modelling volatility as driven by a single multiplicative factor produces a likelihood featuring a variance matrix with Kronecker structure, which in turns ensures the existence of a naturally conjugate N-IW prior (conditional on the history of volatilities).

To derive the likelihood of the VAR, consider the equations for all observations $t = 1, ..., T$. By stacking them by columns and then transposing the system we get:

$$Y = X\Phi + U, \tag{14}$$

where Y is a $T \times N$ data-matrix with rows y_t', X is a $T \times k$ data-matrix with rows $x_t' = (1, y_{t-1}', y_{t-2}', ..., y_{t-p}')$ and U is a $T \times N$ data-matrix with rows u_t'. Now consider the likelihood of the VAR conditional on knowledge of the history of volatilities. Define a diagonal matrix having the whole history of λ_t in the main diagonal:

$$\Lambda = diag(\lambda_1, ..., \lambda_T). \tag{15}$$

In Appendix C we show that the likelihood is given by:

$$\begin{aligned} P(Y|\Phi, V, \Lambda) \propto |V|^{-0.5k} \exp\{-0.5tr[V^{-1}(\Phi - \hat{\Phi})(X'\Lambda^{-1}X)(\Phi - \hat{\Phi})]\} \\ \times |V|^{-0.5(T-k)} \exp\{-0.5tr[V^{-1}\hat{S}]\}, \end{aligned} \tag{16}$$

where $\hat{\Phi} = (X'\Lambda^{-1}X)^{-1}X'\Lambda^{-1}Y$ and $\hat{S} = (Y - X\hat{\Phi})'\Lambda^{-1}(Y - X\hat{\Phi})$. Equation (16) shows that this distribution — conditional on the knowledge of Λ — can be decomposed in an inverse Wishart marginal distribution for V (second line) and a conditional (on V) matricvariate normal distribution for Φ (first line). A distinguishing feature of the matricvariate normal distribution is the Kronecker structure of its covariance matrix.

For such form of likelihood function a naturally conjugate N-IW prior is available. For example, one could use the N-IW Minnesota-type prior proposed by Kadiyala and Karlsson (1997). Here we derive the prior as in Del Negro and Schorfheide (2004), i.e. by using the

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

moments of the underlying state space system. Details of the derivation can be found in Appendix C. The resulting prior distribution is:

$$\Phi|V,\theta,\gamma \sim N(\hat{\Phi}^*(\theta), V \otimes (\gamma T \Gamma_{X^{*\prime}X^*}(\theta))^{-1}), \tag{17}$$

$$V|\theta,\gamma \sim IW(\hat{S}^*(\theta), \gamma T - k), \tag{18}$$

where:

$$\hat{\Phi}^*(\theta) = \Gamma^{-1}_{X^{*\prime}X^*}(\theta)\Gamma_{X^{*\prime}X^*}(\theta), \tag{19}$$

$$\hat{S}^*(\theta) = \gamma T(\Gamma_{Y^{*\prime}Y^*}(\theta) - \Gamma_{Y^{*\prime}X^*}(\theta)\Gamma^{-1}_{X^{*\prime}X^*}(\theta)\Gamma_{X^{*\prime}Y^*}(\theta)). \tag{20}$$

The matrices $\Gamma_{Y^{*\prime}Y^*}(\theta), \Gamma_{Y^{*\prime}X^*}(\theta), \Gamma_{X^{*\prime}X^*}(\theta)$ contain the moments of the yields under the GATSM model and can be easily computed from the state space representation for any given θ using the expressions of $E[y_t y_t']$ and $E[y_t y_{t-h}']$ given by equations (9) and (10).

Note that the prior in (17)-(18) is conditional on θ as the computation of the moment matrices requires knowledge of these parameters. The prior is also conditional on the parameter γ which measures the overall tightness of the prior. Del Negro and Schorfheide (2004) show how this prior can be interpreted as a set of $T^* = \gamma T$ artificial observations obeying the state space (6)-(7), and therefore the tightness parameter γ can be interpreted as the fraction of artificial to actual observations in the sample. In the Bayesian terminology, the prior in (17)-(18) is said to be hierarchical, i.e. dependent on a second layer of parameters (hyperparameters) — θ and γ — for which priors will be specified and posteriors distributions obtained.

The joint posterior distribution of Φ and V, conditional on θ, γ,and Λ, will be proportional to the likelihood times the prior, and by adding the appropriate integrating constant will be of the N-IW form (Zellner, 1973):

$$\Phi|Y,\Lambda,V,\theta,\gamma \sim N(\tilde{\Phi}(\theta), V \otimes (\gamma T \Gamma_{X^{*\prime}X^*}(\theta) + X'\Lambda^{-1}X)^{-1}), \tag{21}$$

$$V|Y,\Lambda,\theta,\gamma \sim IW(\tilde{S}(\theta), (\gamma+1)T - k), \tag{22}$$

where :

$$\tilde{\Phi}(\theta) = (\gamma T \Gamma_{X^{*\prime}X^*}(\theta) + X'\Lambda^{-1}X)^{-1}(\gamma T \Gamma_{X^{*\prime}Y^*}(\theta) + X'\Lambda^{-1}Y), \tag{23}$$

$$\tilde{S}(\theta) = [(\gamma T \Gamma_{Y^{*\prime}Y^*}(\theta) + Y'\Lambda^{-1}Y) - (\gamma T \Gamma_{Y^{*\prime}X^*}(\theta) + Y'\Lambda^{-1}X)(\gamma T \Gamma_{X^{*\prime}X^*}(\theta) + X'\Lambda^{-1}X)^{-1}(\gamma T \Gamma_{X^{*\prime}Y^*}(\theta) + X'\Lambda^{-1}Y)]. \tag{24}$$

When $\gamma \to 0$ the posterior mean of Φ approaches the OLS estimate. On the other hand, when $\gamma \to \infty$, the posterior mean of Φ approaches the prior mean $\Phi^*(\theta)$, i.e. the value consistent with the sharp GATSM restrictions.

The joint p.d.f. of the posterior distribution of the VAR parameters and the hyperparameters, conditional on the history of volatilities Λ, can be factorized as follows:

$$p(\Phi, V, \theta, \gamma | Y, \Lambda) \propto p(\Phi | Y, \Lambda, V, \theta, \gamma) p(V | Y, \Lambda, \theta, \gamma) p(\theta, \gamma | Y, \Lambda). \tag{25}$$

Draws from the distribution of $\Phi, V, \theta, \gamma | Y, \Lambda$ can be obtained by drawing sequentially from $\theta, \gamma | Y, \Lambda$, $V | Y, \Lambda, \theta, \gamma$, and $\Phi | Y, \Lambda, V, \theta, \gamma$. Draws from $\theta, \gamma | Y, \Lambda$ can be obtained using Metropolis steps using the kernel of the p.d.f. of this distribution, which is available and provided in Appendix D, equation (101). Draws from $V | Y, \Lambda, \theta, \gamma$, and $\Phi | Y, \Lambda, V, \theta, \gamma$ are instead obtained via MC steps using (21) and (22). Finally, we note that in (25) we have omitted conditioning on ϕ, i.e. the parameters of the law of motion for the volatility, because they are redundant under knowledge of Λ.

Drawing from the distribution of $\Phi, V, \theta, \gamma | Y, \Lambda$ allows one to use equations (11) and (12) to produce the predictive density of y_t conditional on the volatility at time t. To complete the model, one needs to simulate the volatility process described by (13); therefore we now turn to the joint posterior of the volatility processλ_t and its law of motion parameters ϕ, conditional on the VAR coefficients:

$$p(\Lambda, \phi | Y, \Phi, V). \tag{26}$$

Note we have omitted to condition also on the hyperparameters because under knowledge of Φ and V they do not yield any additional information. Draws from $\Lambda, \phi | Y, \Phi, V$ can be obtained by drawing in turn from $\phi | Y, \Lambda$ and $\Lambda | Y, \Phi, V, \phi$. Following Cogley and Sargent (2005) we specify conjugate priors on the parameters in ϕ, so that the conditional posterior distribution of ϕ is known and draws from it can be obtained via a MC step.

To draw from the conditional posterior of Λ, we use the method proposed in Carriero, Clark, and Marcellino (2012). Such method is a modification of Cogley and Sargent (2005) to allow for a single stochastic volatility factor. Defining the orthogonalized residuals $w_t = (w_{1t}, ..., w_{nt}) = V^{-1/2} u_t$ the kernel of $p(\Lambda | Y, \Phi, V, \phi)$ is given by:

$$p(\Lambda | Y, \Phi, V, \theta, \gamma, \phi) = \prod_{t=1}^{T} p(\lambda_t | \lambda_{t-1}, \lambda_{t+1}, \phi, w_t) \tag{27}$$

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

$$a = \min \left(\frac{\lambda_t^{*-n \times 0.5} \prod_{i=1}^{n} \exp(-0.5 w_{it}^2 / \lambda_t^*)}{\lambda_t^{-n \times 0.5} \prod_{i=1}^{n} \exp(-0.5 w_{it}^2 / \lambda_t)}, 1 \right). \tag{28}$$

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

An algorithm drawing in turn from (25) and (26) will recover the joint posterior of the VAR coefficients, the hyperparameters, and the volatility. For future reference we label this model $JSZ-VAR-CSV$ (VAR with Joslin, Singleton, Zhu prior and Common Stochastic Volatility).

3.1 Homoskedastic version

In our forecasting exercise we will also consider a homoskedastic version of the model, which we label $JSZ-VAR$ (VAR with Joslin, Singleton, Zhu prior). This model is simply obtained by setting $\lambda_t = 1$ for all t, and is given by:

$$y_t = \Phi_0 + \Phi_1 y_{t-1} + ... + \Phi_p y_{t-p} + \ \epsilon_t; \qquad \epsilon_t \sim N(0, V), \tag{29}$$

Note that the model above is similar to that of the $JSZ - VAR - CSV$ in equations (11)-(13), the only difference being that the volatility is assumed to be constant over time, therefore the parameters ϕ and the volatility λ_t drop out of the analysis and the relevant joint posterior distribution of interest is (25). Also, as under this specification $\Lambda = I_T$, we have that the conditional posteriors of the VAR parameters become:

$$\Phi|Y, \Lambda, V, \theta, \gamma \sim N(\tilde{\Phi}(\theta), V \otimes (\gamma T \Gamma_{X^{*\prime}X^*}(\theta) + X'X)^{-1}), \tag{30}$$

$$V|Y, \Lambda, \theta, \gamma \sim IW(\tilde{S}(\theta), (\gamma+1)T - k), \tag{31}$$

where:

$$\tilde{\Phi}(\theta) = (\gamma T \Gamma_{X^{*\prime}X^*}(\theta) + X'X)^{-1}(\gamma T \Gamma_{X^{*\prime}X^*}(\theta) + X'Y), \tag{32}$$

$$\begin{aligned}\tilde{S}(\theta) = [&(\gamma T \Gamma_{Y^{*\prime}Y^*}(\theta) + Y'Y) - (\gamma T \Gamma_{Y^{*\prime}X^*}(\theta) + Y'X)(\gamma T \Gamma_{X^{*\prime}X^*}(\theta) \\ &+ X'X)^{-1}(\gamma T \Gamma_{X^{*\prime}Y^*}(\theta) + X'Y)].\end{aligned} \tag{33}$$

This model is of course nested in the more general $JSZ - VAR - CSV$ model, and is conceptually identical to the model estimated by Carriero (2011). However, note that differently from Carriero (2011), this model is based on the JSZ representation of the GATSM, which makes estimation much easier and the mixing of the MCMC sampler much better.

4 MCMC estimation

We have now developed all the blocks we need to estimate the VAR with common stochastic volatility and no arbitrage prior ($JSZ - VAR - CSV$). The parameters to be estimated are: (i) the deep coefficients of the model θ; (ii) the GATSM prior tightness γ; (iii) the VAR variances V; (iv) the VAR coefficients Φ; (v) the time series of the stochastic volatility factor Λ; and (vi) the coefficients of the volatility process ϕ. The joint and marginal posterior distributions of these coefficients can be obtained by a MCMC sampling scheme. In particular the algorithm will work as follows:

1. Draw from the conditional posterior distribution of γ, $p(\gamma|Y, \theta, \Lambda)$;
2. Draw from the conditional posterior distribution of θ, $p(\theta|Y, \gamma, \Lambda)$;
3. Draw from the conditional posterior distribution of V, $p(V|Y, \theta, \gamma, \Lambda)$;
4. Draw from the conditional posterior distribution of Φ, $p(\Phi|Y, V, \theta, \gamma, \Lambda)$;
5. Draw from the conditional posterior distribution of Λ, $p(\Lambda|Y, \Phi, V, \phi)$; and
6. Draw from the conditional posterior distribution of ϕ, $p(\phi|Y, \Lambda)$.

Note that steps 1-4 allow us to retrieve draws from $\Phi, V, \theta, \gamma|Y, \Lambda, \phi$, while steps 5-6 provide draws from $\Lambda, \phi|Y, \Phi, V, \theta, \gamma$, and therefore cycling through these two groups of steps resembles a Gibbs sampler and provides draws from the joint posterior of $\Phi, V, \theta, \gamma, \Lambda, \phi|Y$.

The priors on the VAR coefficients V and Φ are set up hierarchically, using equations (17) and (18). We do not need a prior on the first observation of the volatility process as in Cogley and Sargent (2005), because in our setup this value is constrained to 1 to achieve identification of the variance matrix of the disturbances $\Sigma_t = \lambda_t V$. Therefore, we only need to specify priors for γ, θ, ϕ. We use a weakly informative prior for θ, implementing the belief that the first factor in the GATSM is a random walk, the second is stationary but very persistent, and the third is moderately persistent. For γ we use a weakly informative normally distributed prior, and implement the restriction $\gamma > (k + N)/T$, necessary for the priors on V and Φ to be proper, by truncating the posterior draws. The prior mean for γ is centered on 1, which corresponds to giving a-priori the same weight to the GATSM and the unrestricted VAR. More details on the prior distributions can be found in Appendix D.

The marginal posteriors $p(\theta|Y, \gamma, \Lambda)$ and $p(\gamma|Y, \theta, \Lambda)$ do not have a known form but can be factorized as $p(\theta|Y, \gamma, \Lambda) \propto p(Y|\theta, \gamma, \Lambda)p(\theta, \gamma, \Lambda)$ and $p(\gamma|Y, \theta, \Lambda) \propto p(Y|\theta, \gamma, \Lambda)p(\theta, \gamma, \Lambda)$, where $p(Y|\theta, \gamma, \Lambda)$ is available (see equation (101) in Appendix D), which opens the way

for drawing from these distributions using Metropolis-Hastings (MH) steps, as e.g. in Del Negro and Schorfheide (2004) and Giannone, Lenza and Primiceri (2012).

Given the draw of deep parameters θ and γ obtained in Step 1 and Step 2, draws from $p(V|Y,\theta,\gamma,\Lambda)$ and $p(\Phi|Y,V,\theta,\gamma,\Lambda)$ in Step 3 and Step 4 can be obtained by a simple MC draw based on the closed form expressions (23) and (24). In Step 5, draws from the conditional posterior distribution of the stochastic volatility factor $p(\Lambda|Y,\Phi,V,\phi)$ are obtained using a (sequence of) Metropolis step(s) based on the density kernel (27) and the acceptance probability (28). Finally, Step 6 draws from $p(\phi|Y,\Lambda)$ using standard results for the univariate linear regression model. More details on the algorithm are provided in Appendix D.

By iterating on steps $1-6$ we get the full joint and marginal posteriors of $\Lambda,\Phi,V,\theta,\gamma,\phi$. Estimation of the homoskedastic version of the model proceeds along the same lines, but of course for this case steps 5 and 6 are not needed.

5 Empirical application

In this section we present an empirical application of our method using U.S. data. Data are zero-coupon Fama-Bliss yields, at monthly frequency, for maturities 3 months and 1, 2, 3, 5, 7, and 10 years and are plotted in Figure 1.[11] Our sample extends from January 1985 through December 2007, which covers a relatively stable monetary policy regime, and is in line with JSZ and Joslin, Priebsch, and Singleton (2012). Indeed, there is substantial evidence that the Federal Reserve changed its policy rule during the early 1980's (Clarida, Gali, and Gertler 2000, Taylor 1999), while the 2008 financial crisis and the consequent non-standard monetary policy led to highly unusual movements in the term structure, and eventually resulted in short-term interest rates getting close to the zero lower bound (Bauer and Rudebusch 2013, Christensen and Rudebusch 2013).

We estimate all of our VAR specifications using 3 lags, chosen via the Bayesian Information Criterion computed on the full sample.

All the results in the paper are based on four parallel MCMC chains. Each chain is composed of 15,000 draws, from which we eliminate the first 2500 as burn-in, and on which we retain each 25-th draw, for a total of 500 clean draws per chain, which provides 2000 clean draws in total when merging the draws from the different chains. As we detail in Appendix D, we initialize the algorithm using the posterior mode of the model, conditional on a maximum likelihood estimate of the common stochastic volatility factor.

[11] We thank Robert Bliss for providing us with the data.

5.1 In-sample results

We start with in–sample estimation of the $JSZ - VAR - CSV$ model using the complete sample period of 1985-2007.

Table 1 contains information about the convergence of the MCMC algorithm. Panel A displays descriptive statistics for the Inefficiency Factors (IFs) proposed by Geweke (1996), while Panel B contains the Potential Scale Reduction Factors (PSRFs) of Gelman and Rubin (1992). The IFs are computed separately for each chain and then pooled together before computing the descriptive statistics, while the PSRFs are based on the computation of the between-chain and within-chain variance of the four independent chains. A value of the IFs below 20 and of the PSRFs below 1.2 are generally taken as indication that each of the chains has converged to the target distribution (see e.g. Justiniano and Primiceri 2008). As is clear from the figures in Table 1, our algorithm shows good mixing properties and has achieved convergence. As it is reasonable to expect, the homoskedastic version of the model converges faster than the heteroskedastic one and shows better mixing properties.

Table 2 contains estimates of the structural parameters of the model. In order to highlight the effect that the tightness parameter γ has on the estimates, besides the results based on the full estimation of the model where γ has been integrated out, we also report estimates obtained by keeping γ fixed to some values, along with results obtained by maximum likelihood estimation of the JSZ model. In the first group of columns the table reports results for the homoskedastic version of the model, the $JSZ - VAR$. When the tightness γ approaches infinity, this model approaches the JSZ model, whose estimation results, based on maximum likelihood estimation, are reported in the last column of the table. Finally, the columns in the middle of the table report results for the $JSZ - VAR - CSV$.

Our estimates are in line with those reported by JSZ[12]. In particular,our estimates of the parameters $\lambda_1^{\mathbb{Q}}$, $\lambda_2^{\mathbb{Q}}$, $\lambda_3^{\mathbb{Q}}$ are -0.0023, -0.034, and -0.148, close to the values estimated by JSZ, -0.0024, -0.0481, and -0.0713. These values imply an almost nonstationary first factor, a highly persistent second factor, and a persistent third factor. Our estimate of $k_\infty^{\mathbb{Q}}$ is 0.032, which corresponds to a value for the long run mean of the short term rate under the risk neutral measure of 13.91 percent (per annum).[13] With regard to the conditional

[12] In comparing with JSZ, we consider the results for the specification they label RPC, which is the closest to ours, with some differences. In particular, the RPC specification of JSZ is based on a model in which the factors are priced without error, and the data-set is slightly different, as we use the 3-month rate rather than the 6-month rate. Finally, note that we report results for $k_\infty^{\mathbb{Q}}$ while JSZ report the value of the long run mean of the short term rate under the risk-neutral measure, which is given by $r_\infty^{\mathbb{Q}} = -k_\infty^{\mathbb{Q}}/\lambda_1^{\mathbb{Q}}$.

[13] This value is well above the value of 8.61 obtained by JSZ in their baseline specification, but is close to some other values they report for some alternative specifications (11.2). This happens because our sample is

variances of the factors, our values are in general smaller than those reported in JSZ, due to the fact that our specification allows for measurement errors which of course implies a better in-sample fit.

The estimates of $JSZ-VAR-CSV$ and $JSZ-VAR$ models are broadly similar. Form the table is clear that, as expected, when the tightness parameter γ increases the posterior means tend to move towards the JSZ estimates, and the posterior variances shrink. For the case in which γ is estimated on the full sample, its mean is 0.48 with a standard deviation of 0.056 for the $JSZ - VAR - CSV$ model, and 0.47 with a standard deviation of 0.053 for the $JSZ - VAR$ model. In the recursive samples which we will use in our forecasting exercise, γ ranges from 0.4 to 1.2.

Figure 2 displays the posterior distribution of the common stochastic volatility factor λ_t, while Figure 3 displays the implied time series of the stochastic volatilities for each of the yields in the VAR, obtained using the estimates of λ_t and equation (12). 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.

5.2 Forecasting exercise

We now consider a pseudo-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 reach the end of the sample, using a last estimation window that includes data up to November 2007.

In the forecast comparisons we consider three models. The first is the $JSZ-VAR-CSV$ model, featuring both time variation in volatility and shrinkage towards the JSZ restrictions. The second model, the $JSZ - VAR$, features only shrinkage towards the JSZ restrictions while the volatilities are kept constant. The third model, labelled $BVAR - CSV$, does feature time varying volatility, but the shrinkage is towards the prior mean and variances of a Minnesota style prior, i.e. it is implementing the a-priori belief that the yields follow univariate random walks.[14] We compute the relevant moment matrices $\Gamma_{\tilde{X}^{*\prime}\tilde{X}^{*}}, \Gamma_{\tilde{Y}^{*\prime}\tilde{X}^{*}}$, and

different both in the time series and in the cross-sectional dimension. In particular we use the 3-month rate, and data from 1985, which explains the higer value of its mean.

[14] The Minnesota-style prior we implement is the same as 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 these priors are in line with the

$\Gamma_{\tilde{Y}^{*\prime}\tilde{Y}^{*}}$ based on this alternative prior and then we use them in expressions (23) and (24). Besides the use of these prior matrices, all the remaining characteristics of the model are left unchanged and the model is estimated using the same MCMC sampler as the JSZ-VAR-CSV. In particular, also the overall tightness on this prior is optimally chosen by estimating the parameter γ via a Metropolis step.

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. 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 forecast interval).

We compare the performance of the considered models against forecasts produced by a simple random walk. Our use of the random walk as a benchmark is based on a large body of evidence documenting that such a model is particularly difficult to beat in term structure forecasting. Several term structure models have a hard time in improving over a simple random walk forecast, especially so at short horizons, as documented in several studies including Duffee (2002), Diebold and Li (2006), Christensen, Diebold and Rudebusch (2011), and Carriero, Kapetanios and Marcellino (2012). Point forecasts from the random walk are simply 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 for yields and innovations to yields with variance equal to the variance of changes in yields over the estimation sample.

belief that macroeconomic data typically feature unit roots, and the "dummy initial observation" 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 e.g. Leeper, Sims, and Zha (1996), Robertson and Tallman (1999), Waggoner and Zha (1999), and Zha (1998), and more recently Giannone, Lenza and Primiceri (2012) and Carriero, Clark and Marcellino (2013).

5.3 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 Error ($RMSFE$) for a given model. Let $\hat{y}_{t+h}^{(i)}(M)$ denote the forecast of the i-th yield $y_{t+h}^{(i)}$ made by model M. The $RMSFE$ made by model M in forecasting the i-th variable at horizon h is:

$$RMSFE_{i,h}^{M} = \sqrt{\frac{1}{P}\sum\left(\hat{y}_{t+h}^{(i)}(M) - y_{t+h}^{(i)}\right)^2}, \tag{34}$$

where the sum is computed over all the P forecasts produced.

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 (2011a,b) 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 don't consider rejections of the alternative model in favor of the benchmark). In the tables we will present, differences in accuracy that are statistically different from zero are denoted by one, two, or three asterisks, corresponding to significance levels of 10%, 5%, and 1%, respectively. The underlying p-values are based on t-statistics computed with a serial correlation-robust variance, using a rectangular kernel, $h-1$ lags, and the small-sample adjustment of Harvey, Leybourne, and Newbold (1997).

The overall calibration of the density forecasts can be measured with the average of log predictive likelihoods (density scores henceforth), motivated and described in, e.g., Geweke and Amisano (2010). For model M_i, the h-step ahead score is defined as:

$$SCORE_{i,h}^{M} = \frac{1}{P}\sum \log p(y_{t+h}^{(i)}|y^{(t)}, M), \tag{35}$$

where the sum is computed over all the P forecasts produced, $y_{t+h}^{(i)}$ denotes the observed outcome for the data in period $t+h$, and $y^{(t)}$ denotes the history of data up to period t (the sample used to estimate the model and form the prediction for period $t+h$). The predictive density $p(\cdot)$ is obtained by univariate kernel estimation based on the MCMC output.

To provide a rough gauge of the statistical significance of differences in density scores, 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 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 expanding, sample of data. 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 don't consider rejections of the alternative model in favor of the benchmark). In the tables we will present, differences in average scores that are statistically different from zero are denoted by one, two, or three asterisks, corresponding to significance levels of 10%, 5%, and 1%. The underlying p-values are based on t-statistics computed with a serial correlation-robust variance, using a rectangular kernel, $h-1$ lags, and the small-sample adjustment of Harvey, Leybourne, and Newbold (1997).

5.4 Out-of-sample forecasting results

We now turn to the evaluation of the point and density forecasting performance of the proposed model. The evaluation is based on forecasts produced for the period going from January 1995 to December 2007, using the recursive scheme described in section 5.2.

Table 3 and Table 4 present results for point and density forecasts, respectively. In the tables, the first panel contains the RMSFEs and SCOREs obtained by using the random walk forecasts, for which the underlying forecast units are basis points. The remaining panels display the relative RMSFEs and differences in SCOREs of the competing models relative to the random walk. A figure below 1 in the relative RMSFEs, or above 0 in the SCOREs, signals that a model is outperforming the random walk benchmark. As the SCOREs are measured in logs, a score difference of say 0.05, signals a 5% gain in terms of density forecast accuracy. The best model for each forecast horizon and yield is highlighted in bold, while the stars *,**,***, to the right of the entries signal rejection of the null of equal forecast accuracy at 10, 5, and 1 percent level.

Several conclusions emerge from the tables. Focusing on point forecasts, both the $JSZ-VAR$ and the $JSZ-VAR-CSV$ in general outperform the random walk benchmark. The gains tend to increase with the forecast horizon, and to be relatively higher for shorter maturities, consistent with the GATSM evidence in such studies as Christensen, Diebold and Rudebusch (2011) and Diebold and Rudebusch (2013).

At short horizons, the $JSZ-VAR$ and the $JSZ-VAR-CSV$ models perform equally, as the effect of changing volatility is small on point forecasts at such short horizons. However, at longer horizons, the presence of stochastic volatility combined with the nonlinearities inherent in multi-step forecasting implies that the forecasts from the two models tend to differ, with the $JSZ-VAR$ model being slightly better, although formal tests of equal accuracy

(not reported in the table) reveal that such differences are never statistically significant.

Arguably, shrinkage can help forecasting regardless of the direction in which it is applied, simply because it reduces the problem of over-parameterization typical of a large Vector Autoregression. However, the $JSZ-VAR$ and $JSZ-VAR-CSV$ models both improve over the BVAR with a Minnesota style prior and common stochastic volatility. This shows that it is not only the use of shrinkage per se, but also the direction of such shrinkage which yields the forecasting gains. Hierarchical modelling of both the prior means and prior variances of a VAR coefficients helps more than a hierarchical modelling limited to the prior variances.

Focussing on density forecasts, the $JSZ-VAR-CSV$ model systematically produces the best density forecasts, outperforming all remaining models, including the random walk benchmark.

As in the case of the point forecasts, the gains against the random walk benchmark are increasing with the forecast horizon, for both the $JSZ-VAR-CSV$ and the $JSZ-VAR$ model, but the contribution of variation in volatility to such forecasting gains is decreasing with the forecast horizon. For example, the 1-step ahead forecast of the 3-month yield made by the $JSZ-VAR-CSV$ obtains a gain vis-a-vis the random walk which is more than double of the gain obtained by the $JSZ-VAR$ (30% vs 13%). The 12-step ahead forecast of the same variable leads higher gains against the random walk, but the scores under the $JSZ-VAR-CSV$ and $JSZ-VAR$ models are much closer (*126*% and 119% respectively). This is due to the fact that, at longer horizons, the projected volatilities tend to converge to their unconditional mean, thereby reducing the difference between homoskedastic and heteroskedastic models. Such a feature is not specific to our model, but is likely to appear in any model featuring (stationary) time variation in the volatilities.

Also, note that the density forecasting performance of the $BVAR-CSV$ is strongly related to its point forecasting performance: the $BVAR-CSV$ produces good/bad density forecasts whenever it produces good/bad point forecasts (at the short/long end of the curve, respectively). This suggests that while the presence of variation in volatility does help this model to produce a reasonable assessment of uncertainty around the point forecasts, the fact that, under the prior, such density forecasts are centered around a random walk forecast rather than a GATSM-based forecast reduces its overall forecasting performance.

To summarize, both the $JSZ-VAR-CSV$ and the $JSZ-VAR$ models produce competitive point and density forecasts, systematically outperforming the RW benchmark. The gains against the random walk increase with the forecast horizon. The $JSZ-VAR-CSV$ specification produced the best density forecasts throughout the sample. The gains

in using a specification with time varying volatility tend to die out as the forecast horizon increases.

5.5 Subsample analysis

In order to assess the stability of our results throughout the sample, we have computed the loss functions (RMSFE and SCORE) recursively. In particular, starting from January 1996, we compute the relative RMSFE and SCORE difference of the $JSZ - VAR - CSV$ against the RW, based on the sample 1995:1 to 1995:12, then we add one more forecast evaluation period and repeat the computation, and so forth until the last evaluation period, i.e. December 2007, is reached.

Results of this exercise are displayed in Figure 4 and Figure 5 for point and density forecast, respectively. In order to avoid cluttering the graphs, we focus only on the four combinations given by the shortest and longest maturity in our sample (3-month and 10-year rate) and by the shortest and longest forecast horizon (1- and 12- step ahead). 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 downwards the forecasting performance of the $JSZ - VAR - CSV$ is improving (relative to the RW), while when it is trending upwards, the forecasting performance is deteriorating. From an inspection of the picture several conclusions can be drawn.

At the 1-step ahead forecast horizon, the $JSZ-VAR-CSV$ is overall outperforming the RW throughout the sample, *but the relative gains were not stable*. For the short term yield, the $JSZ-VAR-CSV$ performs well at the beginning of the sample, but then the forecasting performance deteriorates during the end of the nineties, and by 1999 the performance is on average the same as the one of the RW. From 1999 onwards, the performance steadily improves and reaches the average RMSFE of 0.85%. For the 10-year yield instead, the deterioration in forecasting performance is slower, the relative RMSFE is always below 1 before 2004, and then stays quite steadily around this value until the end of the sample.

At the long forecast horizon the behavior is rather different. In this case, the $JSZ - VAR - CSV$ underperforms the RW at the beginning of the sample, but then it starts dramatically improving around 1999, and the improvements continues steadily for the short end of the curve. For the long end, it is interesting to note that, within a similar pattern of improving relative RMSFE, there are some periods in which the forecasting performance

deteriorates, and these are around 2001-2002 and after 2007, which were both characterized by large instability in the market.

Therefore, with regard to point forecast, the overall pattern is that -over time- the $JSZ - VAR - CSV$ forecasts improved for the long-end of the curve and deteriorated for the short-end, and that long-horizon forecasts of the long-end of the curve are particularly problematic during periods of financial instability.

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. Therefore, a value of the time series above 0 signals that the $JSZ - VAR - CSV$ is outperforming the RW in density forecasts. Also, as the series depicted is a recursive mean, whenever the series is trending upwards the forecasting performance of the $JSZ - VAR - CSV$ is improving (relative to the RW), while when it is trending downwards, the forecasting performance is deteriorating. From an inspection of the picture several conclusions can be drawn.

At the 1-step ahead forecast horizon, the $JSZ - VAR - CSV$ is overall outperforming the RW throughout the sample, with quite stable differences in the SCOREs. However it is also apparent that the model works much better during periods of high volatility, while the RW tends to improve its forecasting performance during calmer periods. At the long forecast horizon the pattern is similar, with the $JSZ - VAR - CSV$ performing particularly well in the first part of the sample, up to and including 2001, and then deteriorating slowly during the period 2002-2005. After 2005, the volatility increases and the $JSZ - VAR - CSV$ improves its performance.

5.6 The role of the factor structure and no arbitrage restrictions

In this section we discuss two remaining relevant issues: i) to what extent imposing the GATSM as a prior improves with respect to imposing it exactly, and ii) to what extent the results are driven by the imposition of the no-arbitrage restrictions on the loadings of the system rather than by the other assumptions implicit in the GATSM, and in particular the assumption of a factor structure for the yields.

In order to address the first issue we have computed forecasts based on the GATSM model described by equations (6) and (7). We estimate the model using Gibbs sampling using the same prior set-up as for the $JSZ - VARs$, after estimating the factors in a preliminary step via principal components.[15] As the GATSM is homoschedastic, we compare

[15] A full Bayesian estimation would require filtering the factors using e.g. the Carter and Kohn (1994) algorithm. We do not pursue this strategy here for simplicity, and because as shown in JSZ the filtered estimates of the factors are almost indistinguishable from the principal components. Arguably this choice

it against the homoschedastic version of the model ($JSZ - VAR$). Results for this experiment are presented in Table 5, which contains the relative RMSFE and average difference in SCORE of the GATSM relative to the $JSZ-VAR$. For point forecasts, the $JSZ-VAR$ outperforms the GATSM in all but two cases. The gains are more pronounced at longer forecast horizons, where they can range between 17% and 28%. For density forecasts, the gains are systematic and always significant, more pronounced at short horizons. An inspection at the density forecasts reveals that the GATSM tends to produce too large predictive density, a feature that stems from the fact that it contains disturbances in both the observation and transition equation, while in the $JSZ-VAR$ these disturbances enter indirectly via the specification of the prior mean and variances of the VAR coefficients, on which the requirement of stationarity imposes a constraint on the posterior draws. The overall picture shows that using the model as a prior significantly improves its forecasts, in line with the results of Carriero (2004).

We now turn to the second issue. As we stressed in Section 2, the GATSM-prior imposes: i) a factor structure for the yields, ii) a particular dynamic structure for the factors under the $\mathbb{P}$ measure (a VAR(1) homoskedastic process), and iii) a set of restrictions on the observation equation of the system. Here we want to evaluate the role of these latter restrictions. In order to do so, we have re-estimated the model using a prior that only implements (i) and (ii) without imposing (iii). This can be done by simply concentrating out, via an OLS regression on the principal component of the yields, the coefficients in the vector $A_p(\Theta_P^{\mathbb{Q}})$ and the matrix $B_p(\Theta_P^{\mathbb{Q}})$ appearing in (7), which implies that these coefficients are no longer a function of the deep parameters $\Theta_P^{\mathbb{Q}}$.[16]

Results of this experiment are displayed in Table 6. The results shown in the table are striking, and clearly indicate that the cross-equation no arbitrage restrictions on the loadings only play a minor role in improving forecast accuracy. In terms of point forecasts, the additional gains of imposing the restrictions is minimal and never significant, although it is important to mention the fact that they are always positive. In terms of density forecasts the pattern is similar, with small, positive gains, but in this case the gains can be occasionally larger (e.g. 6% for the 1-step ahead forecast of the 3-month yield) and

implicitly under-estimates the overall uncertainty present in the model, which in principle could worsen its density forecasting performance. However this is not the case in our application, as it turns out that the reason behind the poor performance of the GATSM is the fact that it tends to produce too wide density forecasts with respect to the JSZ-VAR, and therefore filtering the factors would even accentuate such feature.

[16] Concentrating these coefficients out via OLS estimation is strictly speaking imprecise. However we believe that the additional complications in terms of estimation which would arise by estimating these parameters within the Gibbs sampling algorithm offsets the benefits, and that this strategy is reasonable considering the fact our goal is simply to establish the relative importance of the cross-equation restrictions.

significant. The overall picture leads us to conclude that the role of no-arbitrage restrictions on the loadings is minor; they do not help a lot but they definitely do not harm.

These results are in line with the argument of Duffee (2011a), who argues that since 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 regression-based approach.

Summarizing the findings discussed in this section, we find that the forecasting performance improves mainly because a factor structure is imposed as a prior within a more general VAR representation for the yields. Imposing the factor structure exactly considerably worsen the forecasts. One interpretation of this result is that the more general VAR representation is able to capture more information on the dynamics of the yield curve, which instead gets lost in a factor model with 3 factors. In this sense, the additional information on the yields dynamics picked up by our richer VAR specification could be related to the hidden component of Duffee (2011b).[17] Instead relaxing the cross-equation restrictions on the loadings only marginally worsens the forecasts, in line with Duffee (2011a).

6 Conclusions

In this paper we propose a way to impose a no arbitrage affine term structure model as a prior on a vector autoregression, while allowing also for time variation in the error volatilities. As the volatilities of a panel of yields move closely together, we impose on them a factor structure in which the volatility of each yield is ultimately related to a common stochastic volatility factor, as in Carriero, Clark and Marcellino (2012). To shrink the VAR coefficients towards the values implied by an underlying affine term structure model we use a methodology similar to that put forward by Del Negro and Schorfheide (2004).

The affine model towards which VAR coefficients are shrunk is the new canonical form of no arbitrage models recently introduced by Joslin, Singleton, and Zhu (2011). This new representation presents very important advantages in the computation of the likelihood, because it allows one to disentangle the role played by the structural parameters in the risk neutral and physical measure, which in turn allows one to factorize the likelihood and concentrate out some parameters, reducing dramatically the difficulties typically arising in the search for the global optimum.

We provide the conditional posterior distribution kernels of the model and we propose

[17] Duffee (2011b) shows that a great deal of information on future term structures is contained in a hidden component, which enters the dynamics of the yields and at the same time does not explain the (contemporaneous) cross section.

a MCMC algorithm to perform estimation. While we apply the proposed model to term structure forecasting, this is only one of the possible applications of the method, which can be applied for 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. Our proposed model also generalizes the one of Giannone, Lenza and Primiceri (2012), as it specifies hierarchically not only the prior variances but also the prior means of the VAR coefficients. Our results show that a hierarchical specification for the prior means does help in forecasting, highlighting the fact that not only shrinkage per se, but also the direction towards which shrinkage takes place, can be helpful.

By estimating the model using U.S. data on government bond yields covering the period from January 1985 to December 2007, we provide evidence that both the shrinkage toward the affine term structure model and the use of time variation in the volatilities can produce substantial gains in both point and density forecasts. In particular, shrinkage towards a GATSM model provides better point and density forecasts than a random walk, which is typically a very strong benchmark in forecasting yields. Both the point and density forecast gains are increasing with the forecast horizon. The inclusion of time variation in yields volatilities leads to systematic gains in density forecasts with respect to a homoskedastic model, especially so at short horizons.

Our findings show that the forecasting gains are mainly due to the use of a prior imposing a factor structure on the yield within a more general VAR representation for the yields, rather than imposing the factor structure exactly. Instead, the cross-equation restrictions on the loadings only have a marginal role, in line with Duffee (2011a). One interpretation of these results is that the more general VAR representation is able to capture more information on the dynamics of the yields with respect to a factor model with 3 factors. In this sense, the informational gains achieved by the VAR specification could be related to the hidden components of Duffee (2011b).

Appendix A: derivation of the JSZ representation

To make this paper self-contained, we derive here the JSZ representation of the GATSM. A more rigorous and detailed description can be found in JSZ. The evolution of n risk factors (a n-dimensional state vector) is given by:

$$\Delta S_t = K_{0S}^{\mathbb{P}} + K_{1S}^{\mathbb{P}} S_{t-1} + \Sigma_S \varepsilon_t^{\mathbb{P}} \tag{36}$$

$$\Delta S_t = K_{0S}^{\mathbb{Q}} + K_{1S}^{\mathbb{Q}} S_{t-1} + \Sigma_S \varepsilon_t^{\mathbb{Q}} \tag{37}$$

$$r_t = \rho_{0S} + \rho_{1S} S_t, \tag{38}$$

where $\mathbb{Q}$ and $\mathbb{P}$ denote the risk neutral and physical measures, r_t is the short term rate, $\Sigma_S \Sigma_S'$ is the conditional variance of the states and the errors are i.i.d. Gaussian random variables. The model-implied yield on a zero-coupon bond of maturity τ is an affine function of the state S_t (Duffie and Kan 1996):

$$\widetilde{y}_t(\tau) = A_\tau(\Theta_S^{\mathbb{Q}}) + B_\tau(\Theta_S^{\mathbb{Q}}) S_t \tag{39}$$

where $\Theta_S^{\mathbb{Q}} = \{K_{0S}^{\mathbb{Q}}, K_{1S}^{\mathbb{Q}}, \Sigma_S, \rho_{0S}, \rho_{1S}\}$ and the functions $A_\tau(\Theta_S^{\mathbb{Q}})$ and $B_\tau(\Theta_S^{\mathbb{Q}})$ are computed recursively and satisfy a set of Riccati equations:

$$A_{\tau+1} = A_\tau + K_{0S}^{\mathbb{Q}\prime} B_\tau + 0.5 B_\tau' \Sigma_S \Sigma_S' B_\tau - \rho_{0S} \tag{40}$$

$$B_{\tau+1} = B_\tau + K_{1S}^{\mathbb{Q}\prime} B_\tau - \rho_{1S} \tag{41}$$

with initial conditions $A_0 = B_0 = 0$. Here the use of the symbol $\widetilde{\ }$ highlights that the yields $\widetilde{y}_t$are assumed to be perfectly priced by the model, i.e. they contain no measurement error.

A preliminary result (Joslin 2007) is that any GATSM can be re-parametrized as follows:

$$\begin{aligned} K_{0S}^{\mathbb{Q}} &= (k_\infty^{\mathbb{Q}}, 0, ..., 0) && (42)\\ K_{1S}^{\mathbb{Q}} &= J(\lambda^{\mathbb{Q}}) \text{ (real Jordan form)} && (43)\\ \Sigma_S &= \text{lower triangular} \\ \rho_{0S} &= 0 \\ \rho_{1S} &= i \text{ (vector of ones).} \end{aligned}$$

The $\lambda^{\mathbb{Q}}$ are the (ordered) eigenvalues of $K_{1S}^{\mathbb{Q}}$. Note that in this case knowledge of $k_\infty^{\mathbb{Q}}, \lambda^{\mathbb{Q}}, \Sigma_S$ will be sufficient to compute the loadings so we can write $A(\Theta_S^{\mathbb{Q}}) = A(k_\infty^{\mathbb{Q}}, \lambda^{\mathbb{Q}}, \Sigma_S)$ and $B(\Theta_S^{\mathbb{Q}}) = B(\lambda^{\mathbb{Q}})$.

Now consider n linear combinations of N yields (that is, portfolios), and label them $P_t = W y_t$. Assume for the moment that the portfolios P_t are priced without error. JSZ

show that i) the state vector S_t which is in general unobservable can be replaced by the observable portfolios P_t by means of an invariant transformation, and ii) the Q-distribution of the observable portfolios P_t is entirely characterized by $\Theta_P^{\mathbb{Q}} = \{k_\infty^{\mathbb{Q}}, \lambda^{\mathbb{Q}}, \Sigma_P\}$ where Σ_P is the Cholesky factor of the conditional variance of P_t.[18]

To derive the JSZ rotation we start from getting a measurement equation in terms of the states P_t. Rewrite the measurement equation (39) by stacking by columns the equations for different yields:

$$\underset{N\times 1}{\widetilde{y}_t} = \underset{N\times 1}{A(\Theta_S^{\mathbb{Q}})} + \underset{N\times n}{B(\Theta_S^{\mathbb{Q}})}\ \underset{n\times 1}{S_t} \tag{44}$$

with $\widetilde{y}_t = [\widetilde{y}_t(\tau_1), ..., \widetilde{y}_t(\tau_N)]'$, $A(\Theta_S^{\mathbb{Q}}) = [A_{\tau_1}, ..., A_{\tau_N}]'$, and $B(\Theta_S^{\mathbb{Q}}) = [B'_{\tau_1}, ..., B'_{\tau_N}]'$. By premultiplying (44) by W the measurement equation can be stated as:

$$P_t = A_W + B'_W S_t, \tag{45}$$

where

$$A_W = WA(\Theta_S^{\mathbb{Q}}) \tag{46}$$

and

$$B'_W = WB(\Theta_S^{\mathbb{Q}}). \tag{47}$$

From (45) we can get an expression for S_t:

$$S_t = (B_W^{'})^{-1}(P_t - A_W), \tag{48}$$

and substituting (48) into the measurement equation (44) gives:

$$\widetilde{y}_t = A_p + B_p P_t \tag{49}$$

with:

$$A_p = (I - B(\Theta_S^{\mathbb{Q}})(B_W^{'})^{-1}W)A(\Theta_S^{\mathbb{Q}}), \tag{50}$$

$$B_p = B(\Theta_S^{\mathbb{Q}})(B_W^{'})^{-1}, \tag{51}$$

and:

$$\Sigma_P \Sigma'_P = B_W^{'} \Sigma_S \Sigma'_S B_W. \tag{52}$$

Note that since $B(\Theta_S^{\mathbb{Q}}) = B(\lambda^{\mathbb{Q}})$ and $B_W = WB(\Theta_S^{\mathbb{Q}})$, the matrix Σ_S can be derived under knowledge of $\lambda^{\mathbb{Q}}$ and Σ_P, and in turn knowledge of $k_\infty^{\mathbb{Q}}, \lambda^{\mathbb{Q}}, \Sigma_S$ yields the coefficients in

[18] The parameter k^∞ under Q-stationarity (and if the multiplicity of the first eigenvalue λ_1^Q is $m_1 = 1$) is related to the risk neutral long run mean of the short rate as follows: $k_\infty^Q = -\lambda_1^Q r_\infty^Q$. As a result, it is possible to define equivalently $\Theta_P^Q = r_\infty^Q, \lambda^Q, \Sigma_P$.

$A(\Theta_S^{\mathbb{Q}}) = A(k_\infty^{\mathbb{Q}}, \lambda^{\mathbb{Q}}, \Sigma_S)$. It follows that knowledge of $\Theta_P^{\mathbb{Q}} = k_\infty^{\mathbb{Q}}, \lambda^{\mathbb{Q}}, \Sigma_P$ allows one to compute A_p and B_p. The JSZ canonical form corresponding to the measurement equation (49) is:

$$\Delta P_t = K_{0P}^{\mathbb{P}} + K_{1P}^{\mathbb{P}} P_{t-1} + \Sigma_P \varepsilon_t^{\mathbb{P}} \tag{53}$$

$$\Delta P_t = K_{0P}^{\mathbb{Q}} + K_{1P}^{\mathbb{Q}} P_{t-1} + \Sigma_P \varepsilon_t^{\mathbb{Q}} \tag{54}$$

$$r_t = \rho_{0P} + \rho_{1P} P_t. \tag{55}$$

In their Theorem 1 JSZ state that such form is observationally equivalent to that in (36), (37), and (38) above. The relation between the two representations can be derived using the standard formulas for an invariant transformation and is given by:

$$K_{1P}^{\mathbb{Q}} = B_W J(\lambda^{\mathbb{Q}}) B_W^{-1} \tag{56}$$

$$K_{0P}^{\mathbb{Q}} = k_\infty^{\mathbb{Q}} B_W e_{m_1} - K_{1P}^{\mathbb{Q}} A_W \tag{57}$$

$$\rho_{1P} = (B_W^{-1})' i \tag{58}$$

$$\rho_{0P} = -A_W \rho_{1P}, \tag{59}$$

where e_{m_1} is a vector of zeros except for the entry m_1 which is one (m_1 being the multiplicity of the first eigenvalue λ_1^Q).

Now assume the portfolios (and therefore the yields) are measured with error. In this case we define the observed yields as y_t. The interpretation of the portfolios stays the same: the model based factors are $P_t = W\widetilde{y}_t$ as before, but now these differ from the observed factors $P_t^o = W y_t$ so one needs to filter the unobserved states P_t. To do so, for a given W, and $\Theta_P^{\mathbb{Q}} = \{\lambda^{\mathbb{Q}}, k_\infty^{\mathbb{Q}}, \Sigma_P\}$ it is possible to derive $A(\Theta_S^{\mathbb{Q}})$, $B(\Theta_S^{\mathbb{Q}})$, A_W, and B_W as described above. Then it is possible to compute the transition equation parameters using (56) to (59) and the measurement equation parameters using (50) and (51). The measurement error is given by $\Sigma_y \varepsilon_t^y$. The resulting state space system is:

$$\Delta P_t = K_{0P}^{\mathbb{P}} + K_{1P}^{\mathbb{P}} P_{t-1} + \Sigma_P \varepsilon_t^{\mathbb{P}} \tag{60}$$

$$y_t = A_p + B_p P_t + \Sigma_y \varepsilon_t^y. \tag{61}$$

After concentrating out $K_{0P}^{\mathbb{P}}$ and $K_{1P}^{\mathbb{P}}$ in a preliminary OLS step, the vector of coefficients necessary to write the model in the state space form (60)-(61) is:

$$\theta = (\lambda^{\mathbb{Q}}, k_\infty^{\mathbb{Q}}, \Sigma_P, \Sigma_y). \tag{62}$$

Appendix B: derivation of the moments of the yields under JSZ

For a given draw of θ, it is convenient to compute the moments of the yields under the state space model in (60)-(61) as follows. Rewrite the state equation as:

$$P_t = K_{0P}^{\mathbb{P}} + (K_{1P}^{\mathbb{P}} + I)P_{t-1} + \Sigma_P \varepsilon_t^{\mathbb{P}}. \tag{63}$$

Compute the unconditional mean:

$$\bar{P} = E[P_t] = (I - (K_{1P}^{\mathbb{P}} + I))^{-1} K_{0P}^{P} = -{K_{1P}^{\mathbb{P}}}^{-1} K_{0P}^{\mathbb{P}}, \tag{64}$$

which implies $K_{0P}^{\mathbb{P}} = -K_{1P}^{\mathbb{P}}\bar{P}$. Define the demeaned factors $f_t = P_t - \bar{P}$ and write:

$$f_t = P_t - \bar{P} = -K_{1P}^{\mathbb{P}}\bar{P} + (K_{1P}^{\mathbb{P}} + I)P_{t-1} + \Sigma_P \varepsilon_t^{\mathbb{P}} - \bar{P} = (K_{1P}^{\mathbb{P}} + I)f_{t-1} + \Sigma_P \varepsilon_t^{\mathbb{P}} \tag{65}$$

and:

$$y_t = A_p + B_p P_t + \Sigma_y \varepsilon_t^y = A_p + B_p(f_t + \bar{P}) + \Sigma_y \varepsilon_t^y \tag{66}$$

The moments of y_t implied by the state space system can be computed using (65) and (66) and are:

$$E[y_t y_t'] = (A_p + B_p\bar{P})(A_p + B_p\bar{P})' + B_p \Sigma_f B_p' + \Sigma_y \Sigma_y' \tag{67}$$

and

$$E[y_t y_{t-h}'] = (A_p + B_p\bar{P})(A_p + B_p\bar{P})' + B_p(K_{1P}^{\mathbb{P}} + I)^h \Sigma_f B_p', \tag{68}$$

where $\Sigma_f = E[f_t f_t']$ solves the Lyapunov equation $\Sigma_f = (K_{1P}^{\mathbb{P}} + I)\Sigma_f(K_{1P}^{\mathbb{P}} + I)' + \Sigma_P \Sigma_P'$.

Appendix C: VAR coefficients priors and posteriors

In this appendix we derive the priors and posteriors of the VAR coefficients. Once the data have been rescaled by the history of volatilities, the approach described is the one of Del Negro and Schorfheide (2004), to which we refer the reader for further details.

Likelihood

The VAR under consideration is:

$$\begin{aligned} y_t &= \Phi_0 + \Phi_1 y_{t-1} + ... + \Phi_p y_{t-p} + u_t. && (69)\\ u_t &= \lambda_t^{0.5}\epsilon_t,\ \epsilon_t \sim N(0, V), && (70)\\ \log(\lambda_t) &= \phi_0 + \phi_1 \log(\lambda_{t-1}) + \nu_t,\ \nu_t \sim \text{iid } N(0, \phi_2), && (71) \end{aligned}$$

and can be written in matrix form:

$$Y = X\Phi + U, \tag{72}$$

where Y is a $T \times N$ data-matrix with rows y_t', X is a $T \times k$ data-matrix with rows $x_t' = (1, y_{t-1}', y_{t-2}', ..., y_{t-p}')$ and U is a $T \times N$ data-matrix with rows u_t'. Vectorizing the system gives:

$$y = (I \otimes X)vec(\Phi) + vec(U) \tag{73}$$

where the variance of $u = vec(U)$ is:

$$Q = V \otimes \Lambda. \tag{74}$$

with:

$$\Lambda = diag(\lambda_1, ..., \lambda_T). \tag{75}$$

The likelihood is:

$$p(y|\Phi,V,\Lambda) = 2\pi^{-\frac{TN}{2}}|Q|^{-\frac{1}{2}} \exp(-(y - (I \otimes X)vec(\Phi))'\ Q^{-1}(y - (I \otimes X)vec(\Phi))/2) \tag{76}$$

The matrix Q^{-1} can be written as:

$$\begin{aligned} Q^{-1} &= V^{-1} \otimes \Lambda^{-1} \\ &= (I_n \otimes \Lambda^{-1/2})(V^{-1} \otimes I_T)(I_n \otimes \Lambda^{-1/2}) \\ &= \breve{\Lambda}^{-1/2}(V^{-1} \otimes I_T)\breve{\Lambda}^{-1/2}, \end{aligned} \tag{77}$$

where in the last equality we defined $\breve{\Lambda}^{-1/2} = I_N \otimes \Lambda^{-1/2}$. Also, the determinant $|Q|^{-\frac{1}{2}}$ can be written as:

$$\begin{aligned} |\breve{\Lambda}^{-1/2}(V^{-1} \otimes I_T)\breve{\Lambda}^{-1/2}|^{-\frac{1}{2}} &= (|\breve{\Lambda}^{-1/2}||(V^{-1} \otimes I_T)|\breve{\Lambda}^{-1/2}|)^{-\frac{1}{2}} \\ &= (|\breve{\Lambda}^{-1}||(V^{-1}|^T|I_T|^N))^{-\frac{1}{2}} \\ &= |\Lambda|^{N/2}|V|^{T/2}. \end{aligned} \tag{78}$$

Substituting (77) and (78) into (76) gives:

$$\begin{aligned} p(y|\Phi,V,\Lambda) = 2\pi^{-\frac{TN}{2}}|\Lambda|^{N/2}|V|^{T/2} \quad & \exp(-(y - (I \otimes X)vec(\Phi))'\breve{\Lambda}^{-1/2}(V^{-1} \otimes I_T) \\ & \breve{\Lambda}^{-1/2}(y - (I \otimes X)vec(\Phi))\ /2) \end{aligned} \tag{79}$$

Defining the rescaled residuals $r = \breve{\Lambda}^{-1/2}(y - (I \otimes X)vec(\Phi))$, the cross-product term in (79) can be written as $r'(V^{-1} \otimes I_T)r$. As $r'(V^{-1} \otimes I_T)r$ is a scalar, then we can define $vec(R) = r$ and use $tr(D'A'BC) = tr(A'BCD') = vec(A)'(D \otimes B)vec(C)$ to get:

$$p(y|\Phi,V,\Lambda) = 2\pi^{-\frac{TN}{2}}|\Lambda|^{N/2}|V|^{T/2} \exp(-tr(V^{-1}R'R)/2) \tag{80}$$

Note that $R'R = (Y - X\Phi)'\Lambda^{-1}(Y - X\Phi)$ is the scatter matrix of rescaled residuals.[19] Therefore we can write:

$$\begin{aligned} p(y|\Phi,V,\Lambda) &= 2\pi^{-\frac{TN}{2}}|\Lambda|^{N/2}|V|^{T/2}\exp(-tr(V^{-1}(Y'\Lambda^{-1}Y - \Phi'X'\Lambda^{-1}Y \\ &\quad -Y'\Lambda^{-1}X\Phi + \Phi'X'\Lambda^{-1}X\Phi))) /2). \end{aligned} \tag{82}$$

Defining:

$$\hat{\Phi} = (X'\Lambda^{-1}X)^{-1}X'\Lambda^{-1}Y \tag{83}$$

$$\hat{S} = Y'\Lambda^{-1}Y - (Y'\Lambda^{-1}X)(X'\Lambda^{-1}X)^{-1}(X'\Lambda^{-1}Y), \tag{84}$$

the likelihood can be written as:

$$\begin{aligned} P(Y|\Phi,V,\Lambda) \propto |V|^{-0.5k}\exp\{-0.5tr[V^{-1}(\Phi - \hat{\Phi})'X'\Lambda^{-1}X(\Phi - \hat{\Phi})]\} \\ \times |V|^{-0.5(T-k)}\exp\{-0.5tr[V^{-1}\hat{S}]\}. \end{aligned} \tag{85}$$

Equation (85) shows that — conditionally on Λ — the likelihood can be factorized as the product of an inverse Wishart marginal distribution for Σ (second line) and a conditional (on Σ) matricvariate normal distribution for Ψ (first line).

Priors on VAR coefficients

Consider a sample of $T^* = \gamma T$ artificial observations $y^* = [y_1^*, ..., y_{T^*}^*]'$, where y_t^* is a vector of yields of N different maturities, obeying the GATSM model described by equations (6) and (7). As the GATSM features a Moving Average representation, it can be approximated by a sufficiently rich Vector Autoregression. The VAR representation of the artificial data y_t^* is given by:

$$y_t^* = \Phi_0 + \Phi_1 y_{t-1}^* + ... + \Phi_p y_{t-p}^* + u_t^*. \tag{86}$$

$$u_t^* = \lambda_t^{*0.5}\epsilon_t,\ \epsilon_t \sim N(0,V), \tag{87}$$

$$\lambda_t^* = 1 \text{ for all } t. \tag{88}$$

[19] This follows from:

$$\begin{aligned} vec(R) &= \breve{\Lambda}^{-1/2}(y - (I \otimes X)vec(\Phi)) \\ &= \breve{\Lambda}^{-1/2}vec(Y) - \breve{\Lambda}^{-1/2}(I \otimes X)vec(\Phi) \\ &= (I_n \otimes \Lambda^{-1/2})vec(Y) - (I \otimes \Lambda^{-1/2}X)vec(\Phi) \\ &\Rightarrow R = \Lambda^{-1/2}Y - \Lambda^{-1/2}X\Phi = \Lambda^{-1/2}U \end{aligned} \tag{81}$$

Under the GATSM, the common stochastic volatility factor stays constant at its initial value of 1, and the yields y_t^* would have the moment matrices described in (67) and (68). Also this VAR can be written in matrix form:

$$Y^* = X^*\Phi + U^*, \tag{89}$$

where Y^* is a $T \times N$ data-matrix with rows $y_t^{*\prime}$, X^* is a $T \times k$ data-matrix with rows $x_t^{*\prime} = (1, y_{t-1}^{*\prime}, y_{t-2}^{*\prime}, ..., y_{t-p}^{*\prime})$ and U^* is a $T \times N$ data-matrix with rows $u_t^{*\prime}$. The likelihood of the VAR in (86)-(88) is:

$$\begin{aligned} p(y^*|\Phi,V) &= 2\pi^{-\frac{TN}{2}}|V|^{T/2}\exp(-tr(V^{-1}(Y^{*\prime}Y^* - \Phi X^{*\prime}Y^* \\ &\quad -Y^{*\prime}X^*\Phi + \Phi X^{*\prime}X^*\Phi))) /2), \end{aligned} \tag{90}$$

which does not depend on Λ, as under the GATSM we have $\Lambda = I_T$.

The key idea of the method of Del Negro and Schorfheide is to recognize that the likelihood of the artificial data $p(y^*|\Phi,V)$ carries information about Φ and V contained in the sample of artificial observations. Such information is not present in the actual data and can be interpreted as a prior for the coefficients Φ and V. In principle, one could actually simulate samples of artificial data and use them to augment the VAR in (69)-(71). However this would be undesirable, because the prior moment matrices $X^{*\prime}X^*$, $X^{*\prime}Y^*$, $Y^{*\prime}Y^*$ would be affected by stochastic variation. To remove stochastic variation in the prior Del Negro and Schorfheide (2004) substitute the sample moments $X^{*\prime}X^*$, $X^{*\prime}Y^*$, $Y^{*\prime}Y^*$ with their expected values, given by $T^*\Gamma_{X^{*\prime}X^*}$, $T^*\Gamma_{Y^{*\prime}X^*}(\theta)$, and $T^*\Gamma_{Y^{*\prime}Y^*}$, where, for instance, $\Gamma_{Y^{*\prime}Y^*} = E[y_t^* y_t^{*\prime}]$. Note that the population moments in the matrices Γ can be computed using (67)-(68), conditional on knowledge of θ. This means that no simulation of artificial data is actually needed, as for each draw of the deep parameters θ the corresponding values of the moment matrices under the GATSM are available in closed form. Also, note that the total number of artificial observations T^* is a function of γ and therefore the prior moments also depend hierarchically on this hyperparameter.

Adding an initial flat prior $p(\Phi, V) \propto |V|^{-0.5(N+1)}$ and an appropriate constant of integration (see Del Negro and Schorfheide (2004) for details) one can easily derive from (90) a prior for Φ and V:

$$\Phi|V,\theta,\gamma \sim N(\hat{\Phi}^*(\theta), V \otimes \Gamma^{-1}_{X^{*\prime}X^*}(\theta)), \tag{91}$$

$$V|\theta,\gamma \sim IW(\hat{S}^*(\theta), \gamma T - k), \tag{92}$$

with:

$$\hat{\Phi}^*(\theta) = \Gamma^{-1}_{X^{*\prime}X^*}(\theta)\Gamma_{X^{*\prime}Y^*}(\theta), \tag{93}$$

$$\hat{S}^*(\theta) = \gamma T(\Gamma_{Y^{*\prime}Y^*}(\theta) - \Gamma_{Y^{*\prime}X^*}(\theta)\Gamma^{-1}_{X^{*\prime}X^*}(\theta)\Gamma_{X^{*\prime}Y^*}(\theta)), \tag{94}$$

where we have made explicit the conditioning on the hyperparameters θ and γ, and where conditioning onto Λ is not needed as under the GATSM these parameters are fixed.

Posterior of VAR coefficients

The joint posterior distribution of Φ and V, conditional on Λ, will be proportional to the likelihood (85) times the prior (90) :

$$\begin{aligned} p(\Phi,V|\Lambda,y) &\propto |V|^{T/2}\exp(-tr(V^{-1}(Y^{*\prime}Y^{*}-\Phi X^{*\prime}Y^{*} \\ &\quad -Y^{*\prime}X^{*}\Phi+\Phi X^{*\prime}X^{*}\Phi)))\ /2 \\ &\quad |\Lambda^{*}|^{1/2}|V|^{T/2}\exp(-tr(V^{-1}(Y'\Lambda^{-1}Y-\Phi X'\Lambda^{-1}Y \\ &\quad -Y'\Lambda^{-1}X\Phi+\Phi X'\Lambda^{-1}X\Phi)))\ /2) \\ &\propto |V|^{T/2}\exp(-tr(V^{-1}((Y^{*\prime}Y^{*}+Y'\Lambda^{-1}Y)-\Phi(X^{*\prime}Y^{*}+X'\Lambda^{-1}Y) \\ &\quad -(Y^{*\prime}X^{*}+Y'\Lambda^{-1}X)\Phi+\Phi(X^{*\prime}X^{*}+X'\Lambda^{-1}X)\Phi)))\ /2 \end{aligned} \tag{95}$$

Using population moments of the artificial data yields:

$$\begin{aligned} p(\Phi,V|\Lambda,y) &\propto |V|^{T/2}\exp(-\gamma Ttr(V^{-1}((\Gamma_{Y^{*\prime}Y^{*}}(\theta)+Y'\Lambda^{-1}Y)-\Phi(\Gamma_{X^{*\prime}Y^{*}}(\theta)+X'\Lambda^{-1}Y) \\ &\quad -(\Gamma_{Y^{*\prime}X^{*}}(\theta)+Y'\Lambda^{-1}X)\Phi+\Phi(\Gamma_{X^{*\prime}X^{*}}(\theta)+X'\Lambda^{-1}X)\Phi)))/2, \end{aligned} \tag{96}$$

which is the kernel of a N-IW distribution:

$$\Phi|Y,V,\theta,\gamma,\Lambda\sim N(\tilde{\Phi}(\theta),V\otimes(\gamma T\Gamma_{X^{*\prime}X^{*}}(\theta)+X'\Lambda^{-1}X)^{-1}), \tag{97}$$

$$V|Y,\theta,\gamma,\Lambda\sim IW(\tilde{S}(\theta),(\gamma+1)T-k), \tag{98}$$

with:

$$\tilde{\Phi}(\theta) = (\gamma T\Gamma_{X^{*\prime}X^{*}}(\theta)+X'\Lambda^{-1}X)^{-1}(\gamma T\Gamma_{X^{*\prime}Y^{*}}(\theta)+X'\Lambda^{-1}Y) \tag{99}$$

$$\begin{aligned} \tilde{S}(\theta) &= (\gamma T\Gamma_{Y^{*\prime}Y^{*}}(\theta)+Y'\Lambda^{-1}Y) \\ &\quad -(\gamma T\Gamma_{Y^{*\prime}X^{*}}(\theta)+Y'\Lambda^{-1}X)'(\gamma T\Gamma_{X^{*\prime}X^{*}}(\theta) \\ &\quad +'X'\Lambda^{-1}X)^{-1}(\gamma T\Gamma_{X^{*\prime}Y^{*}}(\theta)+X'\Lambda^{-1}Y) \end{aligned} \tag{100}$$

Appendix D: Details on estimation

Priors

The priors on the VAR coefficients V and Φ are set up hierarchically, using equation (17) and (18). Therefore, we only need to specify priors for γ,θ, ϕ. We do not need a prior on

the first observation of the volatility process λ_1 as in Cogley and Sargent (2005), as in our setup this value is constrained to 1 to achieve identification of the variance matrix of the disturbances $\lambda_t V$.

For the parameter γ, which is measuring the degree with which GATSM-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 1 reflects the belief that the GATSM 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.[20]

For the GATSM structural parameters θ we set either a flat or a weakly informative prior. In particular, for the 3 coefficients in λ^Q we set a normal prior $\lambda_1^Q \sim N(-0.002, 0.001)$, $\lambda_2^Q \sim N(-0.02, 0.01)$, $\lambda_3^Q \sim N(-0.2, 0.1)$. Under these prior means the first factor (element of S_t) is virtually a random walk (it features an autoregressive coefficient of 0.99), the second is stationary but very persistent (with an autoregressive coefficient of 0.98), and the third factor is moderately persistent (with an autoregressive coefficient of 0.8). All draws of λ^Q implying non-stationary behavior are discarded, as well as all those for which the relation $\lambda_1^Q > \lambda_2^Q > \lambda_3^Q$ does not hold.[21] For the coefficients Σ_P we set a normal prior centered on the Principal Components estimates of a vector autoregression of the observable factors. 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.[22] For the remaining coefficients k_∞^Q and Σ_y we set an 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 move away a fair amount from them. However, to check for robustness, we have also computed results for a more diffuse version of our priors. In this specification the prior standard deviation of γ is set to 1, and we use a completely uninformative, flat prior on

[20] The values of γ range from 0.4 to 1.2 troughout the recursive samples, and for the full sample the posterior mean of γ is 0.47.

[21] This condition stems from the fact that the coefficients $\lambda_1^Q, \lambda_2^Q, \lambda_3^Q$ are the ordered eigenvalues of the matrix K_{1S}^Q, see equation (43).

[22] In principle, one should not use likelihood information to calibrate the prior, but doing so for error variances and using an auxiliary model (in our case PC) is standard practice, see e.g. Doan Litterman and Sims (1984), Litterman (1986), Sims (1993), Robertson and Tallman (1999), Sims and Zha (1998), Kadiyala and Karlsson (1997), Banbura, Giannone and Reichlin (2010), Koop (2013), and Carriero, Clark and Marcellino (2013).

all the parameters in θ. The main difference emerging in this setup is that the resulting posterior means of γ are higher. In particular the posterior mean of γ is estimated to be 1.5 instead of 1.1. This is driven by the fact that the more diffuse prior on θ is restricting less the data under the GATSM prior, and therefore the overall level of misspecification of the model decreases, thereby increasing the posterior estimate of γ. However, the overall mass of the posterior of γ is unchanged, and still lies roughly between 0.25 and 3, and the posterior means of all the remaining parameters are very similar. This holds in particular for the estimates of the VAR matrices Φ and Σ_t, which are the ones on which the forecasts are ultimately produced, and as a consequence the results of the forecasting exercise are qualitatively identical under this uninformative version of the prior. However, the uninformative prior does deteriorate the mixing of the algorithm, and convergence of the MCMC scheme under such prior requires one to roughly double the number of replications.

MCMC scheme

A key ingredient of our MCMC procedure is the conditional distribution $p(Y|\theta,\gamma,\Lambda)$, which can be obtained in closed form using the Bayes formula and the integrating constants of prior, likelihood, and posterior:

$$\begin{aligned} p(Y|\theta,\gamma,\Lambda) &= p(Y|\Phi,V,\Lambda)p(\Phi,V|\theta,\gamma,\Lambda)/p(\Phi,V|Y,\Lambda) \\ &= \frac{\left|\gamma T\Gamma_{X^{*\prime}X^{*}}(\theta)+X^{\prime}\Lambda^{-1}X\right|^{-\frac{q}{2}}\left|\tilde{S}(\theta)\right|^{-\frac{(\gamma+1)T-k}{2}}}{\left|\gamma T\Gamma_{X^{*\prime}X^{*}}(\theta)\right|^{-\frac{q}{2}}\left|\tilde{S}^{*}(\theta)\right|^{-\frac{\gamma T-k}{2}}} \\ &\times(2\pi)^{\frac{-qT}{2}}\frac{2^{\frac{q((\gamma+1)T-k)}{2}}\Pi_{i=1}^{q}\Gamma[((\gamma+1)T-k+1-i)/2]}{2^{\frac{q(\gamma T-k)}{2}}\Pi_{i=1}^{q}\Gamma[(\gamma T-k+1-i)/2]}. \end{aligned} \tag{101}$$

In expression (101), $\Gamma[\cdot]$ denotes the gamma function and the second equality comes from the normalization constants of the Normal-Inverted Wishart distributions.

The algorithm for estimation draws in turn from the following conditional posterior distributions:

1. Draw from the conditional posterior distribution of γ, $p(\gamma|Y,\theta,\Lambda)$;
2. Draw from the conditional posterior distribution of θ, $p(\theta|Y,\gamma,\Lambda)$;
3. Draw from the conditional posterior distribution of V, $p(V|Y,\theta,\gamma,\Lambda)$;
4. Draw from the conditional posterior distribution of Φ, $p(\Phi|Y,V,\theta,\gamma,\Lambda)$;

5. Draw from the conditional posterior distribution of Λ, $p(\Lambda|Y, \Phi, V, \phi)$; and

6. Draw from the conditional posterior distribution of ϕ, $p(\phi|Y, \Lambda)$;

Note that steps 1-4 allow to retrieve draws from $\Phi, V, \theta, \gamma|Y, \Lambda, \phi$, while steps 5-6 provide draws from $\Lambda, \phi|Y, \Phi, V, \theta, \gamma$, and therefore cycling through these two groups of steps resembles a Gibbs sampler and provides draws from the joint posterior $\Phi, V, \theta, \gamma, \Lambda, \phi|Y$.

Step 1: Drawing from $\gamma|Y, \theta, \Lambda$

The p.d.f. $p(\gamma|Y, \theta, \Lambda)$ does not have a known form but it can be factorized as $p(\gamma|Y, \theta, \Lambda) \propto p(Y|\theta, \gamma, \Lambda)p(\theta, \gamma, \Lambda)$. The prior distribution $p(\theta, \gamma, \Lambda)$ can be further factorized as $p(\gamma)$ times $p(\theta, \Lambda)$, but the latter is constant in this step. Therefore we have $p(\gamma|Y, \theta, \Lambda) \propto p(Y|\theta, \gamma, \Lambda)p(\gamma)$. As $p(Y|\theta, \gamma, \Lambda)$ is given by (101) and $p(\theta)$ is known, draws from $\gamma|Y, \theta, \Lambda$ can be obtained trough a random walk Metropolis step.

Step 2: Drawing from $\theta|Y, \gamma, \Lambda$

The p.d.f. $p(\theta|Y, \gamma, \Lambda)$ can also be factorized as $p(\theta|Y, \gamma, \Lambda) \propto p(Y|\theta, \gamma, \Lambda)p(\theta, \gamma, \Lambda)$. The prior distribution $p(\theta, \gamma, \Lambda)$ can be further factorized as $p(\theta)$ times $p(\gamma, \Lambda)$, with the latter constant in this step. Therefore we have $p(\theta|Y, \gamma, \Lambda) \propto p(Y|\theta, \gamma, \Lambda)p(\theta)$. As $p(Y|\theta, \gamma, \Lambda)$ is given by (101) and $p(\theta)$ is known, draws from $\theta|Y, \gamma, \Lambda$ can be obtained trough a Metropolis step. To improve the mixing we use a multiple block Metropolis-Hastings algorithm, drawing in turn from three blocks of elements of θ: λ^Q and k_∞^Q, Σ_P, and Σ_y. The candidate draws are generated through random walk steps, with variances calibrated using the Hessian of the model at the posterior mode.

Step 3: Drawing from $V|Y, \theta, \gamma, \Lambda$

Draws from $V|Y, \theta, \Lambda$ can be obtained via a MC step using expressions (98).

Step 4: Drawing from $\Phi|Y, V, \theta, \gamma, \Lambda$

Draws from $\Phi|Y, V, \theta, \Lambda$ can be obtained via a MC step using expressions (97).

Step 5: Drawing from $\Lambda|Y, \Phi, V, \phi$

For a given draw of Φ, V, θ, ϕ we can draw the stochastic volatilities as did Carriero, Clark, and Marcellino (2012). Their method is a modification of Cogley and Sargent (2005) to allow for a single stochastic volatility factor. The kernel of $p(\Lambda|Y, \Phi, V, \theta, \phi)$ is given by:

$$p(\Lambda|Y,\Phi,V,\phi) = \prod_{t=1}^{T} p(\lambda_t|\lambda_{t-1},\lambda_{t+1},\phi,w_t), \tag{102}$$

where $w_t = (w_{1t},...,w_{nt}) = V^{-1/2}u_t$ is a vector of orthogonalized residuals and where the disappearance of all values beyond 1 lead and lag is a consequence of the Markov property of the process assumed for λ_t. As the rescaled residuals w_t contain all the information given by Y, Φ, V, we have substituted conditioning with respect to these variables with conditioning with respect to w_t to simplify the notation. The generic element $p(\lambda_t|\lambda_{t-1},\lambda_{t+1},\phi,w_t)$ in the products (102) can be factorised as:

$$\pi(\lambda_t|\lambda_{t-1},\lambda_{t+1},\phi,w_t) \propto p(w_t|\lambda_t,\phi)p(\lambda_{t+1}|\lambda_t,\phi)p(\lambda_t|\lambda_{t-1},\phi) \tag{103}$$

The p.d.f. $p(w_t|\lambda_t,\phi)$ is a normal, while $p(\lambda_{t+1}|\lambda_t,\phi)$ and $p(\lambda_t|\lambda_{t-1},\phi)$ are log-normal. Writing them down we have:

$$\begin{aligned} p(w_t|\lambda_t,\phi) &\propto \lambda_t^{-0.5}\exp(-0.5w_{1t}^2/\lambda_t) \times \lambda_t^{-0.5}\exp(-0.5w_{2t}^2/\lambda_t) \\ &\quad \times ... \times \lambda_t^{-0.5}\exp(-0.5w_{nt}^2/\lambda_t) \qquad (104) \\ p(\lambda_{t+1}|\lambda_t,\phi)p(\lambda_t|\lambda_{t-1},\phi) &= \lambda_t^{-1}\exp(-0.5(\ln\lambda_t - \mu_t)^2/\sigma_c^2) \qquad (105) \end{aligned}$$

In (104) there are no cross terms because the orthogonalized residuals are by construction independent. Equation (105) comes from the product of the two lognormals, and it is slightly different at the beginning and end of the sample. The parameters μ_t and σ_c^2 are the conditional mean and variance of $\log\lambda_t$ given λ_{t-1} and λ_{t+1}. For periods 2 through $T-1$, the conditional mean and variance are $\mu_t = (\phi_0(1-\phi_1) + \phi_1(\log\lambda_{t-1} + \log\lambda_{t+1}))/(1+\phi_1^2)$ and $\sigma_c^2 = \sqrt{\phi_2/(1+\phi_1^2)}$, respectively (the conditional mean and variance are a bit different for period T, while in period 1 λ_t is set to 1 to achieve identification of the variance matrix of the disturbances $\lambda_t V$).

Draws from λ_t are obtained using a sequence of Metropolis steps starting from $t = 2$**.** and ending in $t = T$. In each period, a candidate draw λ_t^* is extracted from a proposal distribution, and then it is accepted with probability a. By choosing as proposal distribution $q(\theta) \propto p(\lambda_{t+1}|\lambda_t,\phi)p(\lambda_t|\lambda_{t-1},\phi)$ the acceptance probability will simplify to:

$$a = \min\left(\frac{p(w_t|\lambda_t^*,\phi)}{p(w_t|\lambda_t,\phi)}, 1\right); \tag{106}$$

where:

$$\frac{p(w_t|\lambda_t^*,\phi)}{p(w_t|\lambda_t,\phi)} = \frac{\lambda_t^{*-n\times0.5}\prod_{i=1}^{n}\exp(-0.5w_{it}^2/\lambda_t^*)}{\lambda_t^{-n\times0.5}\prod_{i=1}^{n}\exp(-0.5w_{it}^2/\lambda_t)}. \tag{107}$$

Note this differs from Cogley and Sargent (2005), as in their case each volatility process $\lambda_{it},\ i = 1, ...n$, is drawn separately conditional on the remaining $n-1$, which means that $n-1$ elements in the products $\Pi_{i=1}^{n} \exp(-0.5 w_{it}^2/\lambda_t^*)$ and $\Pi_{i=1}^{n} \exp(-0.5 w_{it}^2/\lambda_t)$ would cancel out.

Step 6: Drawing from $\phi|Y, \Lambda$

Finally, given the Normal priors for ϕ_0 and ϕ_1 and the Inverse Gamma prior for ϕ_2, the posterior distributions are conjugate and can be computed using standard methods.

Initialization of the algorithm

We set the initial value for γ to its prior mean of 1, which represents equal belief in the GATSM and the unrestricted VAR model. To initialize the algorithm, we start with estimating the GATSM model using maximum likelihood as in JSZ, which provides us with the ML estimates θ^{ML}. Using θ^{ML} and $\Lambda = I$ as initial conditions we maximize with respect to θ the conditional posterior of the model, given by $p(\theta|Y, \gamma, \Lambda) \propto p(Y|\theta, \gamma, \Lambda)p(\theta)$, where $p(Y|\theta, \gamma, \Lambda)$ is given by (101). For the homoskedastic model, this provides us with the value θ^* which maximizes the posterior, and we use this as initial value for θ. For the heteroskedastic model, two additional steps are needed in order to get an initial value for the volatility, and an initial value for θ for the model with varying volatility. We use θ^* to compute the posterior means of the VAR in (21), and derive the implied residuals. We then use these residuals to perform a quasi-maximum likelihood estimation of Λ (based on a Kalman filtering with the adjustment proposed by Harvey, Ruiz and Shepard (1994)). The resulting maximum likelihood estimates Λ^{ML} are used as initial conditions for the volatility. Finally, using Λ^{ML} and θ^* as initial conditions we maximize again the conditional posterior of the model with respect to θ, which provides us with the value θ^{**} which maximizes the posterior of the model conditional on Λ^{ML}. The resulting estimates θ^{**} are used as initial conditions for the MCMC sampler, while the Hessian at the maximum is stored in order to be used to calibrate the Metropolis steps.

References

[1] Amisano, G., Giacomini, R., 2007. Comparing Density Forecasts via Weighted Likelihood Ratio Tests, *Journal of Business and Economic Statistics*, 25, 177-190.

[2] 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.

[3] 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.

[4] Banbura, M., Giannone, D., Reichlin, L., 2010. Large Bayesian Vector Autoregressions, *Journal of Applied Econometrics* 25, 71-92.

[5] Bauer, M.D., Rudebusch G.D., 2013. Monetary Policy Expectations at the Zero Lower Bound, manuscript, Federal Reserve Bank of San Francisco.

[6] Carriero, A., 2011. Forecasting The Yield Curve Using Priors From No- Arbitrage Affine Term Structure Models, *International Economic Review* 52, 425-459.

[7] Carriero, A., Clark, T.E., Marcellino, M., 2012. Common Drifting Volatility in Large Bayesian VARs, CEPR Working Paper #8894.

[8] Carriero, A., Clark, T.E., Marcellino, M., 2013. Bayesian VARs: Specification Choices and Forecast Accuracy, *Journal of Applied Econometrics*, forthcoming.

[9] Carriero, A., Giacomini, R., 2011. How Useful Are No-Arbitrage Restrictions to Forecast the Term Structure of Interest Rates? *Journal of Econometrics* 164, 21-34.

[10] Carriero, A., Kapetanios, G., Marcellino M., 2012. Forecasting Government Bond Yields with Large Bayesian VARs, *Journal of Banking and Finance* 36, 2026-2047.

[11] Carter, C.K., Kohn, R., 1994. On Gibbs Sampling for State Space Models, *Biometrika* 81, 541-553.

[12] 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.

[13] Christensen, J.H.E., Rudebusch, G.D., 2013. Estimating Shadow-Rate Term Structure Models with Near-Zero Yields, manuscript, Federal Reserve of San Francisco.

[14] Clarida, R., Gali, J., Gertler, M., 2000 . Monetary Policy Rules and Macroeconomic Stability: Evidence and Some Theory, *Quarterly Journal of Economics* 115, 147-180.

[15] Clark, T.E, McCracken, M.W., 2011a. Nested Forecast Model Comparisons: A New Approach to Testing Equal Accuracy, manuscript, Federal Reserve Bank of St. Louis.

[16] Clark, T.E., McCracken, M.W., 2011b. Testing for Unconditional Predictive Ability, in *Oxford Handbook of Economic Forecasting*, Michael P. Clements and David F. Hendry, eds., Oxford: Oxford University Press.

[17] Cogley, T., Sargent, T., 2005. Drifts and Volatilities: Monetary Policies and Outcomes in the post-WWII US, *Review of Economic Dynamics* 8, 262-302.

[18] Dai, Q., Singleton, K., 2000. Specification Analysis of Affine Term Structure Models, *Journal of Finance* 55, 1943-1978.

[19] Del Negro, M., Schorfheide, F., 2004. Priors from General Equilibrium Models for VARs, *International Economic Review* 45, 643-673.

[20] Diebold, F.X., Li, C., 2006. Forecasting the Term Structure of Government Bond Yields, *Journal of Econometrics* 130, 337-364.

[21] Diebold, F.X, Mariano, R., 1995. Comparing Predictive Accuracy, *Journal of Business and Economic Statistics* 13, 253-63.

[22] Diebold, F.X., Rudebusch, G.D., 2013. *Yield Curve Modeling and Forecasting*, Princeton University Press, Princeton, New Jersey.

[23] Doan, T., Litterman, R., Sims, C., 1984. Forecasting and Conditional Projection Using Realistic Prior Distributions, *Econometric Reviews* 3, 1-100.

[24] Duffee, G., 2002. Term Premia and Interest Rate Forecasts in Affine Models, *Journal of Finance* 57, 405-443.

[25] Duffee, G., 2009. Forecasting with the Term Structure: The Role of No-Arbitrage, manuscript, University of California-Berkeley.

[26] Duffee, G., 2011a. Forecasting with the Term Structure: The Role of No-Arbitrage Restrictions, Working paper, Johns Hopkins.

[27] Duffee, G., 2011b. Information in (and not in) the Term Structure. *Review of Financial Studies* 24, 2895-2934.

[28] Duffee, G., Stanton, R., 2012. Estimation of Dynamic Term Structure Models, *Quarterly Journal of Finance*, forthcoming.

[29] Duffie, D., Kan, R., 1996. A Yield-Factor Model of Interest Rates, *Mathematical Finance* 6, 379-406.

[30] Gelman, A., Rubin, D.B., 1992. Inference from Iterative Simulation Using Multiple Sequences, *Statistical Science* 7, 457-472.

[31] Geweke, J., 1996. Monte Carlo Simulation and Numerical Integration, in: H.M. Amman, D.A. Kendrick, and J. Rust, *Handbook of Computational Economics* 1, 731-800.

[32] Geweke, J., Amisano, G., 2010. Comparing and Evaluating Bayesian Predictive Distributions of Asset Returns, *International Journal of Forecasting* 26, 16-230.

[33] Giacomini, R., Ragusa, G., 2011. Incorporating Theoretical Restrictions into Forecasting by Projection Methods, manuscript.

[34] Giannone, D., Lenza, M., Primiceri, G., 2012. Prior Selection for Vector Autoregressions, NBER Working Paper No. 18467.

[35] Hamilton, J., Wu, J.C., 2012. Identification and Estimation of Gaussian Affine Term Structure Models, *Journal of Econometrics* 168, 315-331.

[36] Harvey, D., Leybourne, S., Newbold, P., 1997. Testing the Equality of Prediction Mean Squared Errors, *International Journal of Forecasting* 13, 281-291.

[37] Harvey, A., Ruiz, E., Shephard, N., 1994. Multivariate Stochastic Variance Models, *Review of Economic Studies* 61, 247-264.

[38] 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.

[39] Joslin, S., 2007. Pricing and Hedging Volatility in Fixed Income Markets. Mimeo, MIT.

[40] Joslin, S., Singleton, K.J., Zhu, H., 2011. A New Perspective on Gaussian Dynamic Term Structure Models, *Review of Financial Studies* 24, 926-970.

[41] Joslin, S., Priebsch, M., Singleton, K.J., 2012. Risk Premiums in Dynamic Term Structure Models with Unspanned Macro Risks, manuscript.

[42] Justiniano, A., Primiceri, G.E., 2008. The Time Varying Volatility of Macroeconomic Fluctuations, *American Economic Review* 98, 604-641.

[43] Kadiyala, K.R., Karlsson, S., 1997. Numerical Methods for Estimation and Inference in Bayesian VAR-Models, *Journal of Applied Econometrics* 12, 99-132.

[44] Koop, G.M., 2013. Forecasting with Medium and Large Bayesian VARs, *Journal of Applied Econometrics* 28, 177-203.

[45] Leeper, E.M., Sims, C., Zha, T., 1996. What Does Monetary Policy Do? *Brookings Papers on Economic Activity* 27,1-78.

[46] Litterman, R., 1986. Forecasting with Bayesian Vector Autoregressions-Five Years of Experience, *Journal of Business and Economic Statistics* 4, 25-38.

[47] Nelson, C.R., Siegel, A.F., 1987. Parsimonious Modeling of Yield Curve, *Journal of Business* 60, 473-489.

[48] Robertson, J.C., Tallman, E.W., 1999. Vector Autoregressions: Forecasting and Reality, Federal Reserve Bank of Atlanta *Economic Review*.

[49] Robertson, J.C., Tallman, E.W., Whiteman, C.H., 2005. Forecasting Using Relative Entropy, *Journal of Money, Credit and Banking* 37, 383-401.

[50] Shin M. and Zhong M. 2013. Incorporating realized volatility into a dynamic factor model: An application to forecasting bond yield distributions. Mimeo.

[51] Sims, C., 1993. A Nine-Variable Probabilistic Macroeconomic Forecasting Model, in J.H. Stock and M.W. Watson, eds., *Business Cycles, Indicators and Forecasting*, University of Chicago Press, pp. 179-204.

[52] Sims, C., Zha, T., 1998. Bayesian Methods for Dynamic Multivariate Models, *International Economic Review* 39, 949-68.

[53] Taylor, J., 1999. A Historical Analysis of Monetary Policy Rules, in J. Taylor (Ed.), *Monetary Policy Rules*, University of Chicago Press.

[54] Vasicek, O., 1977. An Equilibrium Characterization of the Term Structure, *Journal of Financial Economics* 5, 177-188.

[55] Waggoner DF, Zha T. 1999. Conditional forecasts in dynamic multivariate models. *The Review of Economics and Statistics* 81(4), 639-651

[56] Zellner, A., 1973. *An Introduction to Bayesian Inference in Econometrics*. Wiley.

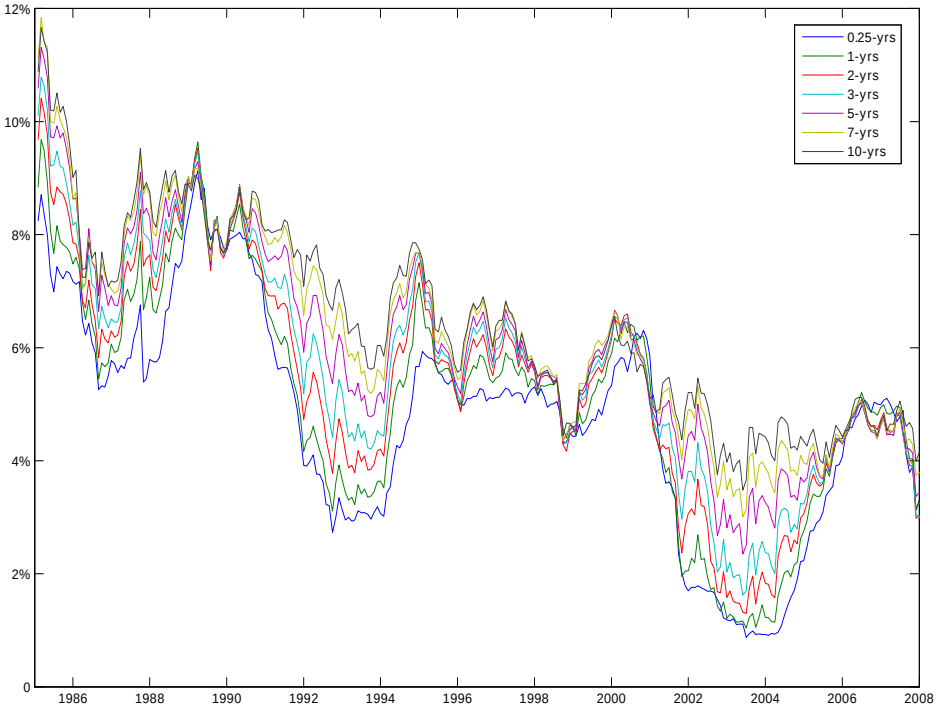

Figure 1: Fama-Bliss zero coupon yields for the US.

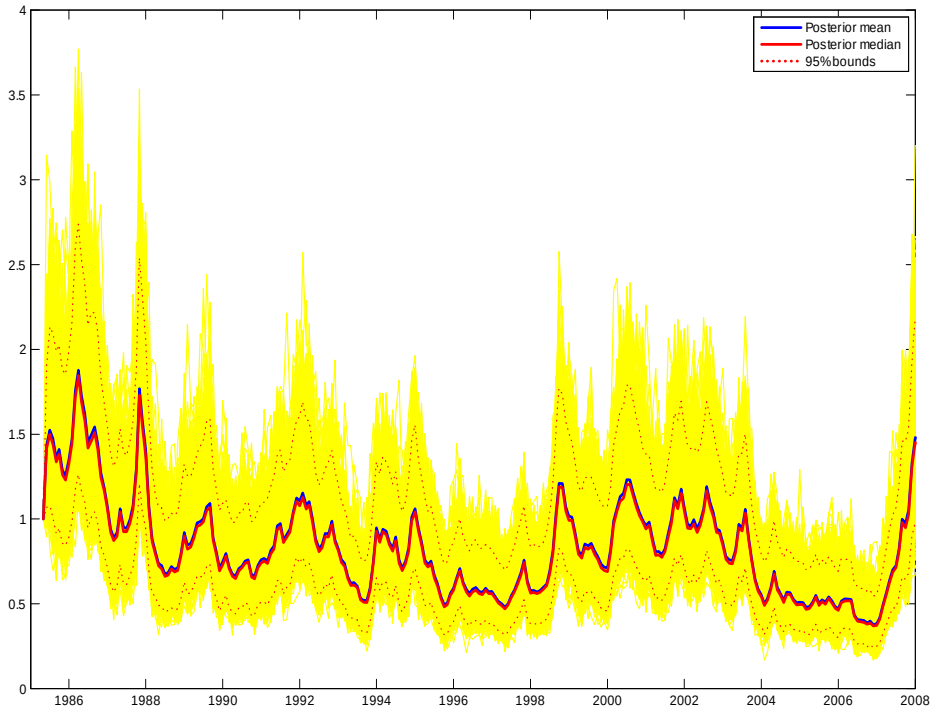

Figure 2: Posterior distribution of the Common Stochastic Volatility process λ_t. The shaded area contains all the final 2000 draws.

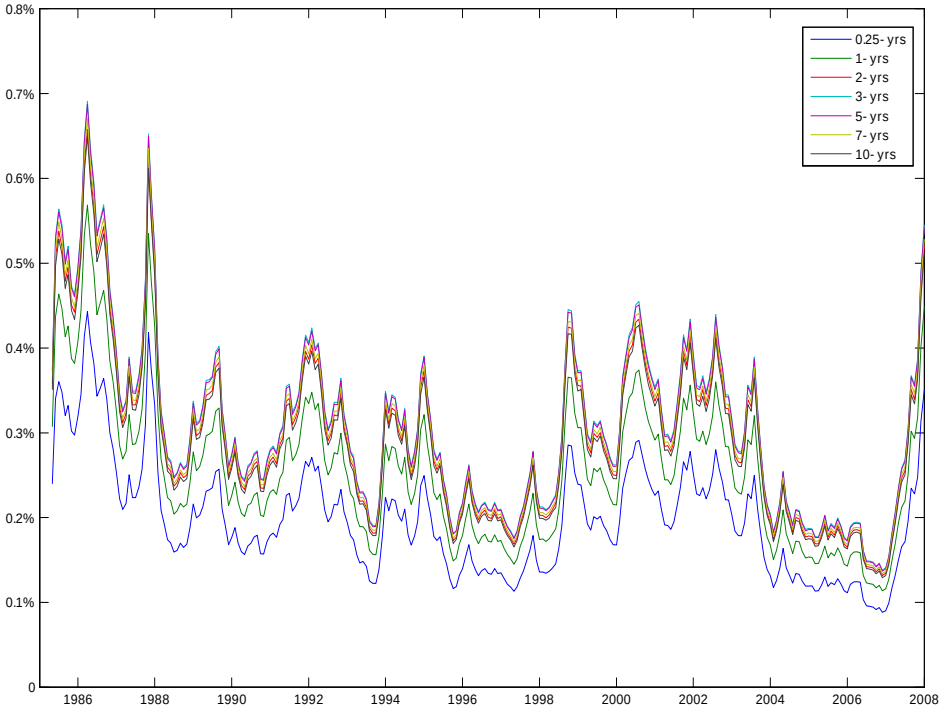

Figure 3: Volatilities for each yield (posterior medians).

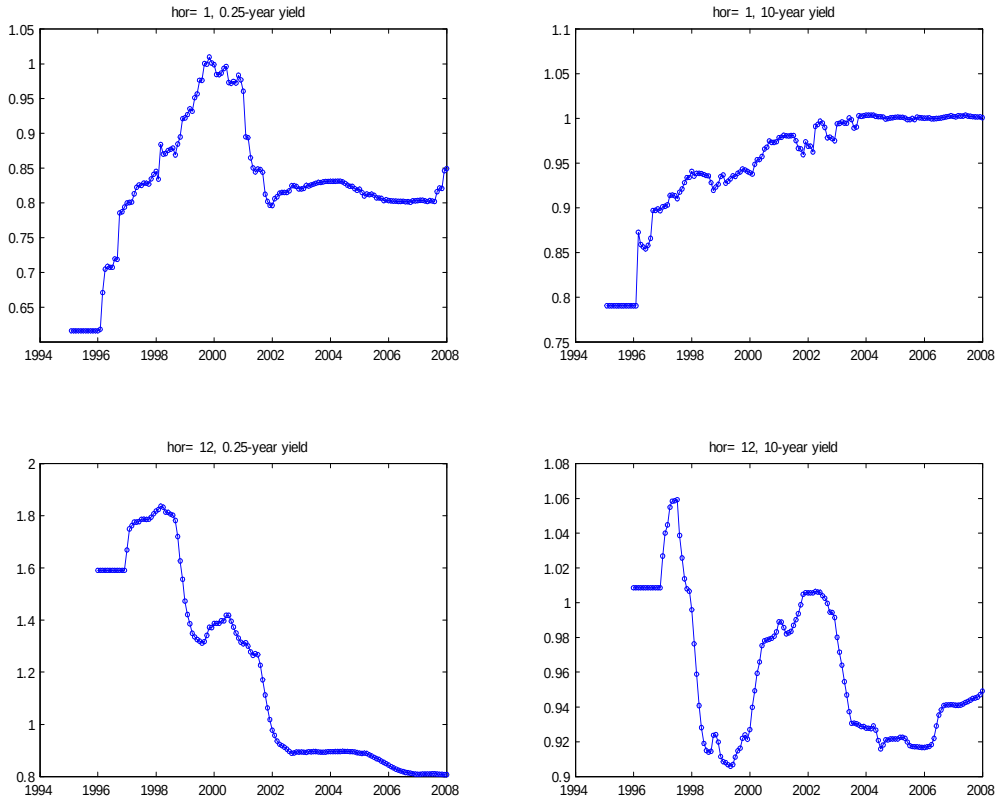

Figure 4: Recursive Relative RMSFE of the $JSZ-VAR-CSV$ versus the Random Walk. Recursive means are computed starting from January 1996.

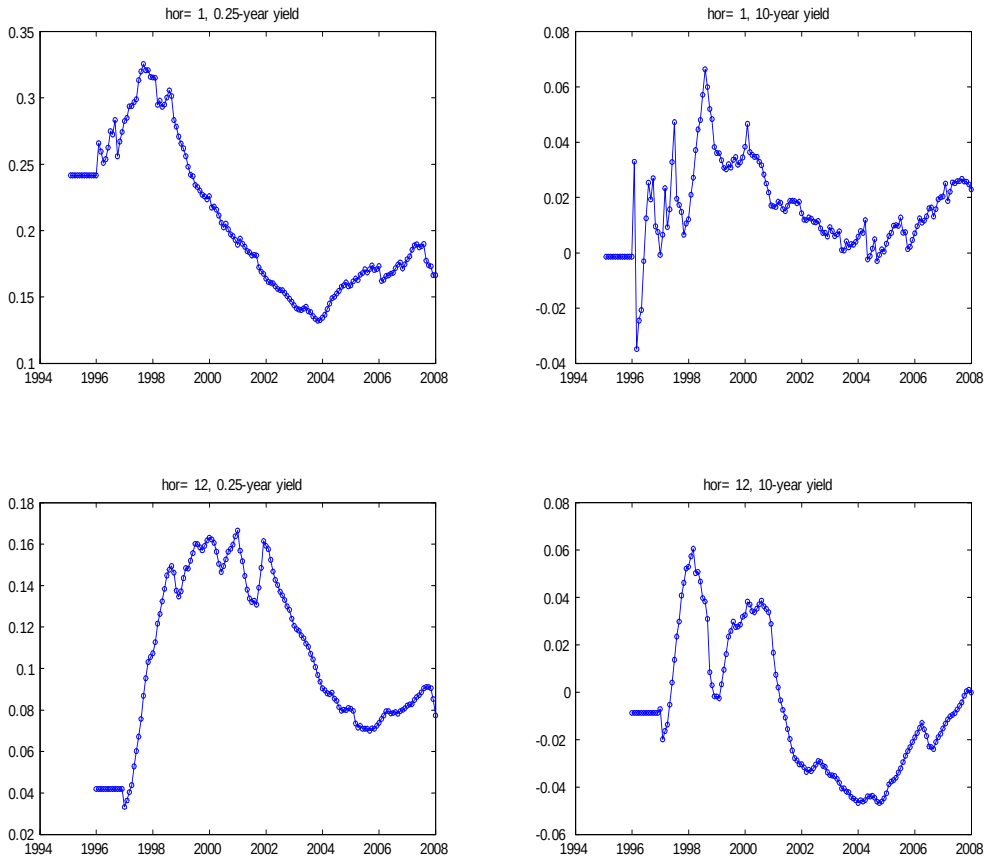

Figure 5: Recursive difference in SCORE of the $JSZ - VAR - CSV$ versus the Random Walk. Recursive differences are computed starting from January 1996.

Table 1. Convergence Diagnostics

	PANEL A: Inefficiency Factors						PANEL B: Potential Scale Reduction Factors					
	Median	**Std**	**5%**	**95%**	**Min**	**Max**	**Median**	**Std**	**5%**	**95%**	**Min**	**Max**
JSZ-VAR												
θ	1.6	0.53	0.84	2.6	0.78	2.6	1	0.001	0.999	1	0.999	1
γ	1	0.11	0.94	1.2	0.94	1.2	0.998	0.005	0.997	1.01	0.997	1.01
V	1.2	0.16	0.95	1.5	0.86	1.6	1	0.002	0.999	1	0.999	1
Φ	0.97	0.23	0.63	1.3	0.46	1.8	1	0.001	0.999	1	0.999	1.01
JSZ-VAR-CSV												
θ	5.1	3.4	1	11	0.77	13	1.07	0.048	1	1.13	0.999	1.13
γ	1.1	0.4	0.6	1.4	0.6	1.4	1	0.004	0.997	1	0.997	1
V	11	2	8.4	15	6.8	15	1.16	0.009	1.14	1.17	1.13	1.17
Φ	0.94	0.22	0.64	1.4	0.46	1.8	1	0.001	0.999	1	0.999	1
Λ	4.7	1.6	2.7	7.9	2	10	1.06	0.012	1.04	1.08	1.03	1.09
φ	2	0.94	1.4	4.6	1.4	4.8	1.02	0.010	1.02	1.03	1.02	1.03

All the results in the paper are based on four parallel MCMC chains. Each chain is composed of 15000 draws, from which we eliminate the first 2500 as burn-in, and on which we retain each 25-th draw, for a total of 500 clean draws per chain, which provides 2000 clean draws in total when merging the draws from the different chains. The IFs are computed separately for each chain and then pooled together before computing the descriptive statistics, while the PSRFs are based on the computation of the between-chain and within-chain variance of the four independent chains.

Table 2. Structural Coefficient Estimates

	JSZ-VAR (homoskedastic)					JSZ-VAR-CSV					JSZ
Tightness:	**0.5**	**1**	**2**	**10**	**integrated out**	**0.5**	**1**	**2**	**10**	**integrated out**	**(inf)**
Parameter (x 100)											
λ_1^Q	-0.224	-0.206	-0.182	-0.236	-0.227	-0.228	-0.220	-0.205	-0.274	-0.229	-0.231
	0.087	0.081	0.073	0.045	0.087	0.089	0.080	0.070	0.042	0.087	0.007
λ_2^Q	-2.855	-2.898	-3.017	-3.638	-2.800	-3.147	-3.251	-3.355	-3.512	-3.118	-3.411
	0.530	0.450	0.390	0.250	0.568	0.537	0.458	0.393	0.198	0.541	0.118
λ_3^Q	-11.990	-11.310	-10.970	-13.610	-12.255	-12.120	-11.544	-11.454	-14.179	-12.091	-14.811
	2.200	1.700	1.400	1.300	2.261	2.008	1.553	1.277	1.118	2.052	0.536
k^Q	0.009	0.009	0.010	0.022	0.009	0.010	0.012	0.015	0.031	0.011	0.032
	0.006	0.006	0.005	0.003	0.006	0.006	0.006	0.005	0.004	0.007	0.001
$\Sigma_{P(1,1)}$	0.674	0.681	0.693	0.716	0.673	0.733	0.786	0.765	0.862	0.775	1.005
	0.048	0.042	0.035	0.032	0.051	0.093	0.098	0.086	0.101	0.099	0.015
$\Sigma_{P(2,1)}$	-0.116	-0.115	-0.115	-0.105	-0.117	-0.159	-0.168	-0.161	-0.155	-0.165	-0.165
	0.021	0.019	0.017	0.014	0.022	0.027	0.026	0.024	0.024	0.028	0.007
$\Sigma_{P(2,2)}$	0.202	0.207	0.211	0.218	0.201	0.212	0.229	0.223	0.254	0.224	0.233
	0.017	0.013	0.012	0.010	0.017	0.030	0.032	0.026	0.030	0.033	0.011
$\Sigma_{P(3,1)}$	-0.071	-0.069	-0.069	-0.066	-0.071	-0.092	-0.097	-0.093	-0.095	-0.095	-0.021
	0.011	0.010	0.009	0.008	0.012	0.015	0.015	0.013	0.014	0.016	0.004
$\Sigma_{P(3,2)}$	0.031	0.033	0.034	0.042	0.031	0.029	0.031	0.033	0.050	0.029	0.045
	0.009	0.008	0.007	0.007	0.009	0.010	0.009	0.008	0.008	0.011	0.008
$\Sigma_{P(3,3)}$	0.089	0.089	0.090	0.097	0.089	0.097	0.102	0.100	0.119	0.103	0.118
	0.009	0.008	0.007	0.006	0.009	0.015	0.016	0.013	0.016	0.017	0.005
$\Sigma_{y(1,1)}$	0.045	0.049	0.051	0.027	0.044	0.060	0.071	0.067	0.035	0.065	2.3E-06
	0.021	0.019	0.018	0.016	0.020	0.023	0.022	0.019	0.021	0.025	8.6E-07
$\Sigma_{y(2,2)}$	0.060	0.064	0.070	0.091	0.060	0.067	0.075	0.078	0.110	0.071	0.111
	0.006	0.005	0.005	0.005	0.006	0.010	0.011	0.010	0.014	0.011	0.005
$\Sigma_{y(3,3)}$	0.027	0.029	0.031	0.034	0.027	0.028	0.031	0.032	0.039	0.030	0.025
	0.005	0.004	0.004	0.005	0.005	0.006	0.006	0.005	0.007	0.007	0.005
$\Sigma_{y(4,4)}$	0.029	0.030	0.032	0.041	0.029	0.028	0.031	0.031	0.043	0.030	0.057
	0.004	0.003	0.003	0.004	0.004	0.005	0.005	0.005	0.006	0.006	0.003
$\Sigma_{y(5,5)}$	0.027	0.029	0.032	0.044	0.027	0.028	0.032	0.033	0.045	0.030	0.068
	0.003	0.003	0.003	0.003	0.003	0.005	0.005	0.005	0.006	0.005	0.004
$\Sigma_{y(6,6)}$	0.037	0.039	0.041	0.050	0.037	0.038	0.041	0.042	0.055	0.040	0.055
	0.004	0.004	0.004	0.005	0.005	0.007	0.007	0.006	0.007	0.007	0.005
$\Sigma_{y(7,7)}$	0.052	0.055	0.059	0.085	0.052	0.058	0.065	0.067	0.106	0.062	0.107
	0.006	0.005	0.005	0.006	0.006	0.010	0.010	0.009	0.014	0.010	0.006

Estimates of the structural coefficients of the model. The entries are posterior means and standard deviations (the figures in smaller size) computed from the MCMC output, except for the last column, where estimates are obtained via maximum likelihood as in JSZ.

Table 3. Evaluation of Point Forecasts. Sample 1995:2007

Maturity→	0.25-yrs	1-yrs	2-yrs	3-yrs	5-yrs	7-yrs	10-yrs
step-ahead ↓							
	RW point forecasting performance						
1	20.80	23.34	27.30	29.02	28.75	27.44	26.32
2	33.71	36.73	42.08	43.66	42.61	40.49	38.31
3	45.22	48.67	53.26	53.52	51.17	48.04	44.67
6	77.98	79.20	80.21	77.21	72.45	66.48	60.29
12	135.78	132.78	122.24	111.25	97.55	86.92	77.52
	JSZ-VAR vs Random Walk						
1	0.86 ***	0.97	1.03	1.01	1.02	1.03	**1.00**
2	**0.80** **	0.95	1.03	1.03	1.03	1.03	**0.99**
3	**0.78** **	0.92	**1.01**	**1.01**	**1.02**	**1.02**	**0.98**
6	**0.78** *	**0.90**	**0.96**	**0.96**	**0.96**	**0.96**	**0.93**
12	**0.80**	**0.85**	**0.86**	**0.86**	**0.86** *	**0.88**	**0.86** *
	JSZ-VAR-CSV vs Random Walk						
1	**0.85** ***	**0.95**	1.02	**1.00**	**1.01**	1.02	**1.00**
2	**0.79** **	**0.93**	1.03	**1.02**	**1.02**	1.03	1.00
3	**0.77** **	**0.91**	**1.01**	**1.01**	**1.02**	**1.02**	1.00
6	**0.78** *	0.91	0.98	0.99	1.00	1.01	0.99
12	0.81 *	0.88	0.91	0.91	0.92	0.94	0.95
	BVAR-CSV vs Random Walk						
1	0.93 ***	0.98	**1.00**	1.01	1.01	**1.01**	1.01
2	0.92 ***	0.99	**1.02**	1.02	1.02	**1.02**	1.02
3	0.92 ***	0.99	1.03	1.03	1.03	**1.02**	1.02
6	0.95 *	1.02	1.06	1.06	1.06	1.04	1.03
12	0.95	1.01	1.06	1.08	1.09	1.08	1.05

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 that a model is outperforming the random walk benchmark. Figures in bold denote that the best model (within the VAR class) for each variable and forecast horizon. Gains in accuracy that are statistically different from zero are denoted by *,**,***, corresponding to significance levels of 10%, 5% and 1% respectively, evaluated using the Diebold and Mariano (1995) t-statistics computed with a serial correlation-robust variance, using a rectangular kernel, h-1 lags, and the small-sample adjustment of Harvey, Leybourne, and Newbold (1997).

Table 4. Evaluation of Density Forecasts. Sample 1995:2007

Maturity→	0.25-yrs	1-yrs	2-yrs	3-yrs	5-yrs	7-yrs	10-yrs
step-ahead ↓							
	RW density forecasting performance						
1	-4.54	-4.66	-4.78	-4.82	-4.81	-4.77	-4.72
2	-4.96	-5.06	-5.17	-5.21	-5.19	-5.15	-5.10
3	-5.26	-5.34	-5.41	-5.41	-5.38	-5.33	-5.26
6	-5.86	-5.81	-5.81	-5.77	-5.73	-5.66	-5.58
12	-7.34	-6.42	-6.24	-6.14	-6.05	-5.97	-5.89
	JSZ-VAR vs Random Walk						
1	0.13 ***	0.05 **	0.00	0.00	0.00	-0.02	-0.02
2	0.16 **	0.03	-0.04	-0.04	-0.03	-0.03	**0.00**
3	0.20 *	0.06	-0.02	-0.03	-0.01	-0.02	**0.01**
6	0.28	0.08	0.03	0.02	0.04	0.02	0.03
12	1.19	0.23	0.10	0.09	**0.10** *	**0.08** *	**0.07** *
	JSZ-VAR-CSV vs Random Walk						
1	**0.30** ***	**0.16** ***	**0.04**	**0.03**	**0.01**	**0.01**	**0.01**
2	**0.29** ***	**0.11** **	**0.00**	**-0.01**	**-0.02**	-0.03	-0.01
3	**0.31** ***	**0.13** **	**0.01**	**0.01**	**0.01**	**-0.01**	0.00
6	**0.37** *	**0.15** *	**0.07**	**0.05**	**0.06**	**0.04**	**0.04**
12	**1.26**	**0.29**	**0.12** *	**0.09**	0.09 *	**0.08** *	**0.07** *
	BVAR-CSV vs Random Walk						
1	0.19 ***	0.12 ***	**0.04**	0.02	0.00	**0.01**	-0.01
2	0.16 ***	0.07 **	**0.00**	-0.02	-0.02	**-0.01**	-0.01
3	0.14 ***	0.07 *	-0.02	-0.03	-0.01	-0.01	-0.01
6	0.12	0.04	0.00	-0.02	0.01	0.01	0.01
12	1.02	0.14	0.00	-0.03	-0.02	0.00	0.02

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 e.g. 0.05 signals a 5% gain in terms of density forecast accuracy. Figures in bold denote that the best model (within the VAR class) for each variable and forecast horizon. Gains in accuracy that are statistically different from zero are denoted by *,**,***, corresponding to significance levels of 10%, 5% and 1% respectively, evaluated using the Amisano and Giacomini (2007) t-statistics computed with a serial correlation-robust variance, using a rectangular kernel, h-1 lags, and the small-sample adjustment of Harvey, Leybourne, and Newbold (1997).

Table 5: JSZ-VAR vs GATSM

Maturity→ step-ahead ↓	0.25-yrs	1-yrs	2-yrs	3-yrs	5-yrs	7-yrs	10-yrs
Relative RMSFE (point forecasting performance)							
1	0.95	0.87**	0.97	1.01	0.98	0.97	0.93**
2	0.92	0.89	0.98	1.01	0.98	0.98	0.95
3	0.88	0.86*	0.94	0.97	0.94	0.93	0.90*
6	0.85	0.84*	0.88	0.90	0.87	0.86	0.81**
12	0.83	0.81*	0.81*	0.80*	0.77*	0.77*	0.72**
Average Difference in SCORE (density forecasting performance) * **							
1	0.640	0.581	0.500	0.493	0.528	0.549	0.602
2	0.588	0.499	0.387	0.385	0.440	0.485	0.553
3	0.536	0.459	0.388	0.404	0.471	0.515	0.596
6	0.408	0.395	0.385	0.422	0.485	0.530	0.603
12	0.245	0.285	0.338	0.405	0.494	0.536	0.592

***** All differences in density forecasts are significant at the 1% level**

The first panel contains the relative RMSFE between the JSZ-VAR and the GATSM exactly imposed. The second panel contains the average difference in SCORE. Gains in accuracy that are statistically different from zero are denoted by *,**,***, corresponding to significance levels of 10%, 5% and 1% respectively, evaluated using the Diebold and Mariano (2005) t-statistics computed with a serial correlation-robust variance, using a rectangular kernel, h-1 lags, and the small-sample adjustment of Harvey, Leybourne, and Newbold (1997). For density forecasts all differences are statistically significant at the 1% level according to the Amisano and Giacomini (2007) t-statistics computed with a serial correlation-robust variance, using a rectangular kernel, h-1 lags, and the small-sample adjustment of Harvey, Leybourne, and Newbold (1997).

Table 6: JSZ-VAR-CSV vs VAR-CSV with factor structure only

Maturity→	0.25-yrs	1-yrs	2-yrs	3-yrs	5-yrs	7-yrs	10-yrs
step-ahead ↓							
Relative RMSFE (point forecasting performance)							
1	1.00	0.99	1.00	1.00	0.99	1.00	1.00
2	1.00	1.00	0.99	0.99	0.99	1.00	1.00
3	0.99	0.99	0.99	0.99	0.99	1.00	1.00
6	0.99	0.99	0.99	0.99	0.99	1.00	1.00
12	0.99	1.00	0.99	0.99	0.99	1.00	1.00
Average Difference in SCORE (density forecasting performance)							
1	0.0664*	0.0166**	0.007	0.005	0.005	0.003	0.009
2	0.0233*	0.007	0.011	0.012	0.005	0.002	0.001
3	0.017	0.003	0.007	0.012	0.013	0.008	0.004
6	0.024	0.007	0.004	0.003	0.001	0.000	0.004
12	0.031	0.0247*	0.022	0.013	0.012	0.009	0.010

The first panel contains the relative RMSFE between the JSZ-VAR-CSV and the same model where the no-arbitrage restrictions on the loadings have not been imposed. The second panel contains the average difference in SCORE. Gains in accuracy that are statistically different from zero are denoted by *,**,***, corresponding to significance levels of 10%, 5% and 1% respectively, evaluated using either the Diebold and Mariano (2005) or the the Amisano and Giacomini (2007) t-statistics computed with a serial correlation-robust variance, using a rectangular kernel, h-1 lags, and the small-sample adjustment of Harvey, Leybourne, and Newbold (1997).