Online Appendix to "Restrictions on Risk Prices in Dynamic Term Structure Models"

Michael D. Bauer*

March 7, 2016

*Federal Reserve Bank of San Francisco, michael.bauer@sf.frb.org

A Change of measure

To show what kind of process the term structure factors follow under $\mathbb{Q}$, I derive the conditional Laplace transform of X_{t+1} under $\mathbb{Q}$. I define the one-period stochastic discount factor (pricing kernel) as

$$M_{t+1} = \exp\left(-r_t - \frac{1}{2}\lambda_t'\lambda_t - \lambda_t'\varepsilon_{t+1}\right).$$

For any one-period pricing kernel the change of measure is implied by

$$M_{t+1} = \exp(-r_t) f^{\mathbb{Q}}(X_{t+1}|X_t)/f^{\mathbb{P}}(X_{t+1}|X_t).$$

Note that the Radon-Nikodym derivative, which relates the densities under the physical and risk-neutral measure, is given by

$$\frac{f^{\mathbb{P}}(X_{t+1}|X_t)}{f^{\mathbb{Q}}(X_{t+1}|X_t)} = \left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right)(X_{t+1};\lambda_t) = \exp\left(\frac{1}{2}\lambda_t'\lambda_t + \lambda_t'\varepsilon_{t+1}\right).$$

I obtain for the risk-neutral conditional Laplace transform

$$\begin{aligned}
E^{\mathbb{Q}}(\exp(u'X_{t+1})|X_t) &= \int \exp(u'X_{t+1}) f^{\mathbb{Q}}(X_{t+1}|X_t) dX_{t+1} \\
&= \int \exp\left(u'X_{t+1} - \frac{1}{2}\lambda_t'\lambda_t - \lambda_t'\varepsilon_{t+1}\right) f^{\mathbb{P}}(X_{t+1}|X_t) dX_{t+1} \\
&= E\left[\exp\left(u'(\mu + \Phi X_t + \Sigma\varepsilon_{t+1}) - \frac{1}{2}\lambda_t'\lambda_t - \lambda_t'\varepsilon_{t+1}\right)|X_t\right] \\
&= \exp\left[u'(\mu - \Sigma\lambda_t + \Phi X_t) + \frac{1}{2}u'\Sigma\Sigma'u\right]
\end{aligned}$$

which is recognized as the conditional moment-generating function of a multivariate normal distribution with mean $\mu - \Sigma\lambda_t + \Phi X_t = (\mu - \lambda_0) + (\Phi - \lambda_1)X_t$ and variance $\Sigma\Sigma'$.

The physical innovations ε_t, which are a vector martingale-difference sequence (m.d.s.) under $\mathbb{P}$, are related to the innovations under $\mathbb{Q}$ by

$$\varepsilon_t^{\mathbb{Q}} = \varepsilon_t + \lambda_{t-1}. \tag{1}$$

Note that the risk-neutral innovations, while being m.d.s. under $\mathbb{Q}$, can have non-zero mean and be predictable under $\mathbb{P}$, depending on the risk price specification.

B Affine bond pricing

Under the assumptions of Section 2.1, model-implied bond prices are exponentially affine functions of the pricing factors:

$$\hat{P}_t^m = e^{\mathcal{A}_m + \mathcal{B}_m X_t},$$

and the loadings $\mathcal{A}_m = \mathcal{A}_m(\mu^{\mathbb{Q}}, \phi^{\mathbb{Q}}, \delta_0, \delta_1, \Sigma)$ and $\mathcal{B}_m = \mathcal{B}_m(\phi^{\mathbb{Q}}, \delta_1)$ follow the recursions

$$\begin{aligned} \mathcal{A}_{m+1} &= \mathcal{A}_m + (\mu^{\mathbb{Q}})'\mathcal{B}_m + \frac{1}{2}\mathcal{B}_m'\Sigma\Sigma'\mathcal{B}_m - \delta_0 \\ \mathcal{B}_{m+1} &= (\phi^{\mathbb{Q}})'\mathcal{B}_m - \delta_1 \end{aligned}$$

with starting values $\mathcal{A}_0 = 0$ and $\mathcal{B}_0 = 0$. Model-implied yields are determined by $\hat{y}_t^m = -m^{-1}\log P_t^m = A_m + B_m X_t$, with $A_m = -m^{-1}\mathcal{A}_m$ and $B_m = -m^{-1}\mathcal{B}_m$.

Risk-neutral yields, the yields that would prevail if investors were risk-neutral, can be calculated using

$$\tilde{y}_t^m = \tilde{A}_m + \tilde{B}_m X_t, \quad \tilde{A}_m = -m^{-1}\mathcal{A}_m(\mu, \phi, \delta_0, \delta_1, \Sigma), \quad \tilde{B}_m = -m^{-1}\mathcal{B}_m(\phi, \delta_1).$$

Risk-neutral yields reflect policy expectations over the life of the bond, $m^{-1}\sum_{h=0}^{m-1} E_t r_{t+h}$, plus a convexity term. The yield term premium is defined as the difference between actual and risk-neutral yields, $ytp_t^m = y_t^m - \tilde{y}_t^m$.

C Risk prices as parameters in a restricted VAR

Estimation of λ, for given values of all other parameters, amounts to estimation of the VAR in (1) subject to the linear constraints in (7). This section lays out the specifics: how to obtain the least squares estimates of λ, which maximize the likelihood for given values of the other parameters (C.1), how to carry out Bayesian inference about λ using the exact conditional posterior (C.2), and how to specify a g-prior in the RVAR context (C.3).

An alternative to RVAR estimation for obtaining estimates of λ is often possible. Subtracting $\mathbb{Q}$-measure expectations from both sides of the VAR in equation (1), the dynamic system becomes

$$X_t - E_{t-1}^{\mathbb{Q}} X_t = \lambda_0 + \lambda_1 X_{t-1} + \Sigma\varepsilon_t,$$

which is a system of Seemingly Unrelated Regressions (SUR) (Zellner, 1962). If the DTSM is estimated using frequentist methods, this formulation can be used to concentrate out λ_0 and λ_1 from the likelihood function—this approach is used, for example, by Joslin et al. (2014). If

the estimation is Bayesian, the conditional posterior for λ can be easily derived (see Zellner and Ando, 2010, and the references therein). For those cases where all rows of (λ_0, λ_1) have at least one non-zero element, the SUR approach is identical to the restricted-VAR approach. However, when one or more rows contain only zeros, the SUR approach is not applicable, because it effectively ignores the equation where all parameters are restricted to zero, whereas the restricted-VAR approach correctly accounts for the inter-equation dependencies of the residuals.

C.1 Likelihood function and least squares estimates

Lütkepohl (2006) describes frequentist estimation of restricted VARs (Section 5.2). Here, I show how inference about λ in a DTSM maps into this context. The VAR in equation (1) can be written in full-data matrix notation as $X = BZ + U$, where $X = (X_1, \ldots, X_T)$, $Z = (Z_0, \ldots, Z_{T-1})$, $Z_t = (1, X_t')'$, $B = (\mu, \Phi)$, and $U = (u_1, \ldots, u_T)$, $u_t = \Sigma\varepsilon_t$. The linear constraints are

$$\beta := vec(B) = \lambda + r = S\lambda_\gamma + r,$$

where S is a known $N(N+1) \times a$ matrix of zeros and ones, λ_γ is an a-vector with the unrestricted elements of λ (according to γ), and $r = vec(\mu^{\mathbb{Q}}, \Phi^{\mathbb{Q}})$. To clarify: λ_γ contains the a unrestricted risk prices, λ contains these as well as $N(N+1) - a$ zeros, and S transforms one into the other ($\lambda = S\lambda_\gamma$). If there are no zero restrictions on λ—as in the case of a maximally-flexible DTSM, and for the SSVS model selection approach—we have $S = I_{N(N+1)}$, $\lambda = \lambda_\gamma$, and there are $a = N(N+1)$ risk prices to estimate.

Plugging in the restrictions and vectorizing the VAR equation we have

$$\begin{aligned} x := vec(X) &= vec(BZ) + vec(U) \\ &= (Z' \otimes I_N)\beta + u \\ &= (Z' \otimes I_N)(S\lambda_\gamma + r) + u \\ z := x - (Z' \otimes I_N)r &= (Z' \otimes I_N)S\lambda_\gamma + u. \end{aligned}$$

The likelihood for z is $N((Z' \otimes I_N)S\lambda_\gamma, I_T \otimes \Omega)$, with $\Omega = \Sigma\Sigma'$. This likelihood is maximized at the generalized least-squares estimate (see Lütkepohl, 2006, eq. 5.2.6) which is given by

$$\hat{\lambda}_\gamma = \left[S'(ZZ' \otimes \Omega^{-1})S\right]^{-1} S'(Z \otimes \Omega^{-1})z. \tag{2}$$

Its estimated covariance matrix is

$$\hat{V}_\gamma = \left[S'(ZZ' \otimes \Omega^{-1})S\right]^{-1}. \tag{3}$$

C.2 Bayesian inference in a restricted VAR

If we assume a natural conjugate prior for λ, which is a normal prior denoted by $N(\underline{\lambda}_\gamma, \underline{V}_\gamma)$, the conditional posterior can be derived in closed form:

$$\begin{aligned} P(\lambda_\gamma|z,\ldots) &\propto \exp\left\{-.5\left[(\lambda_\gamma - \underline{\lambda}_\gamma)'\underline{V}_\gamma^{-1}(\lambda_\gamma - \underline{\lambda}_\gamma)\right.\right. \\ &\quad \left.\left.+(z - (Z' \otimes I_N)S\lambda_\gamma)'(I_T \otimes \Omega)^{-1}(z - (Z' \otimes I_N)S\lambda_\gamma)\right]\right\} \\ &\propto \exp\{-.5[\lambda_\gamma'\underline{V}^{-1}\lambda_\gamma - 2\underline{\lambda}_\gamma\underline{V}^{-1}\lambda_\gamma + \lambda_\gamma'S'(ZZ' \otimes \Omega^{-1})S\lambda_\gamma - 2z'(Z' \otimes \Omega^{-1})S\lambda_\gamma]\} \\ &\propto \exp\left\{-.5\left[\lambda_\gamma'(\underline{V}^{-1} + S'(ZZ' \otimes \Omega^{-1})S)\lambda - 2(\underline{\lambda}_\gamma'\underline{V}^{-1} + z'(Z' \otimes \Omega^{-1})S)\lambda_\gamma\right]\right\}. \end{aligned}$$

The last expression is the Kernel of a Normal distribution with covariance matrix

$$\overline{V}_\gamma = (\underline{V}_\gamma^{-1} + S'(ZZ' \otimes \Omega^{-1})S)^{-1}$$

and mean

$$\overline{\lambda}_\gamma = \overline{V}_\gamma(\underline{V}_\gamma^{-1}\underline{\lambda}_\gamma + S'(Z \otimes \Omega^{-1})z).$$

The (conditional) posterior mean can also be written as

$$\overline{\lambda}_\gamma = \overline{V}_\gamma\left\{\underline{V}_\gamma^{-1}\underline{\lambda}_\gamma + \hat{V}_\gamma^{-1}\hat{\lambda}_\gamma\right\}.$$

These results will be used for drawing λ from its conditional posterior distribution, i.e., using a Gibbs step, in the various MCMC samplers used in this paper.

C.3 g-prior for a restricted VAR

A g-prior for the parameters of a restricted VAR has covariance matrix

$$\underline{V} = g\left[S'(ZZ' \otimes \Omega^{-1})S\right]^{-1} = g\hat{V}$$

so that the posterior mean becomes

$$\overline{\lambda} = \frac{1}{1+g}\underline{\lambda} + \frac{g}{1+g}\hat{\lambda}$$

and the posterior covariance is

$$\overline{V} = \frac{g}{1+g}\left[S'(ZZ' \otimes \Omega^{-1})S\right]^{-1}.$$

The prior used for λ in this paper is an "orthogonalized g-prior" with zero mean. In particular, I calculate $[S'(ZZ' \otimes \Omega^{-1})S]^{-1}$ for the model without restrictions on λ—in which case $S = I_{N(N+1)}$—at the MLE estimates of the remaining parameters (the $\mathbb{Q}$-parameters are needed to calculate Z, and Σ is needed to calculate Ω), and use the diagonal values of the resulting matrix, multiplied by g, as the prior variances.

D Drawing the model parameters

This section provides details on how to draw each block of parameters in the MCMC algorithms. For λ, the approach described here is only used for estimation of a specific model (i.e., conditional on a fixed value of γ), while in the model selection samplers λ and γ are drawn jointly as described in Section E. The other parameters—$k_\infty^{\mathbb{Q}}, \phi^{\mathbb{Q}}, \Sigma$, and σ_e—are always drawn as described here, both in the estimation of a specific model and in the model selection samplers.

D.1 Drawing λ

Conditional on γ and all other model parameters, estimation of the risk-price parameters λ simply amounts to estimating the parameters of an RVAR. This is very convenient, because the conditional posterior for λ is available in closed form, and λ can be sampled using a Gibbs step. The posterior is normal with mean and covariance matrix given in Section C.2. It is a key advantage of the MCMC sampler proposed here that the majority of parameters can be sampled very efficiently using a straightforward Gibbs step.

D.2 Drawing $k_\infty^{\mathbb{Q}}$ and $\phi^{\mathbb{Q}}$

The parameters $k_\infty^{\mathbb{Q}}$ and $\phi^{\mathbb{Q}}$ (the eigenvalues of $\Phi^{\mathbb{Q}}$) exhibit high posterior correlation, due to the fact that they jointly determine the cross-sectional behavior of interest rates. The "grouping-by-correlation" principle for designing efficient MCMC algorithms therefore suggests to draw them jointly as one block. A generally efficient way to do so is to use an Independence MH step with mean and covariance that are close to those of the conditional posterior. Fortunately, such moments can easily be obtained in this context, due to the pa-

rameterization of the term structure model. We can simply take the ML estimates of these parameters (for the unrestricted model) as the mean of the proposal density, as these estimates are generally close to the mode of the conditional posterior. To determine the covariance matrix I obtain a numerical approximation to the Hessian of the conditional posterior at these values and take the negative of its inverse. It turns out to be sufficient to do this only once at the beginning of the algorithm.[1] For the proposal distribution I use a multivariate Student-t distribution with five degrees of freedom. To tailor the proposal density for an Independence MH step in this way is inspired by the approach of Chib and Greenberg (1994), Chib and Ergashev (2009), and Chib and Ramamurthy (2010). In particular, this avoids the need to tune scaling parameters, automatically leads to acceptance probabilities in the 20-50% range that is recommended in the MCMC literature (Gamerman and Lopes, 2006, p. 196), and leads to high efficiency of the sampler (i.e., low autocorrelations of successive draws). Since my approach avoids the need to find the mode of the conditional posterior using numerical optimization (such as the Simulated Annealing algorithm used by Chib and Ramamurthy, 2010), it has very low computational cost.

Because it is useful to have the eigenvalues in $\phi^{\mathbb{Q}}$ sorted and to have similar scaling of parameters that are drawn jointly, I reparameterize them as $\chi = (1000 \cdot k_\infty^{\mathbb{Q}}, \phi_1^{\mathbb{Q}} - 1, \phi_2^{\mathbb{Q}} - \phi_1^{\mathbb{Q}}, \phi_3^{\mathbb{Q}} - \phi_2^{\mathbb{Q}})'$. Due to the prior restrictions on $\phi^{\mathbb{Q}}$, the second through fourth elements of χ are required to lie between -1 and 0. The proposed values are denoted by χ^*, and the proposal density as $q(\cdot)$. The acceptance probability is

$$\alpha(\chi^{(j-1)}, \chi^*) = \min\left\{\frac{P(Y|\chi^*, \theta_-, \gamma)P(\chi^*, \theta_-)q(\chi^{(j-1)})}{P(Y|\chi^{(j-1)}, \theta_-, \gamma)P(\chi^{(j-1)}, \theta_-)q(\chi^*)}, 1\right\},$$

where θ_- denotes all parameters other than $k_\infty^{\mathbb{Q}}$ and $\phi^{\mathbb{Q}}$.

D.3 Drawing Σ

To draw Σ I use a similar Independence MH step as for $(k_\infty^{\mathbb{Q}}, \phi^{\mathbb{Q}})$. Using the operator vech$(\cdot)$ to vectorize the lower triangular of a matrix, the proposal distribution for vech(Σ) is a multivariate t-distribution with five degrees of freedom. The mean is equal to vech(Σ^{MLE}), and the covariance matrix is equal to the negative inverse of the Hessian of the conditional posterior for vech(Σ) at the MLE in the first iteration of the algorithm. After drawing a proposal Σ^*

[1] If the approximate Hessian turns out not to be positive definite, it can be replaced by a positive definite matrix that is close to it, using an eigenvalue decomposition and replacing the negative eigenvalue(s) with small positive eigenvalues.

from this distribution, the acceptance probability is calculated as

$$\alpha(\Sigma^{(j-1)}, \Sigma^*) = \min\left\{\frac{P(Y|\Sigma^*, \theta_-)P(\Sigma^*, \theta_-)q(\text{vech}(\Sigma^{(j-1)}))}{P(Y|\Sigma^{(j-1)}, \theta_-)P(\Sigma^{(j-1)}, \theta_-)q(\text{vech}(\Sigma^*))}, 1\right\},$$

where θ_- denotes all model parameters other than Σ, $q(\cdot)$ is the multivariate-t proposal density, and the prior for Σ only requires that $\Sigma\Sigma'$ is a valid (i.e., positive semi-definite) covariance matrix.

D.4 Drawing σ_e^2

The measurement error variance can be drawn using a Gibbs step, because conditional on the other parameters and the data, the measurement errors can be viewed as regression residuals. For an inverse-gamma (IG) prior with shape parameter $\alpha_0/2$ and scale parameter $\delta_0/2$, the conditional posterior is IG with shape and scale determined by $\alpha_1 = \alpha_0 + n$ and $\delta_1 = \delta_0 + ssr$, where n is the number of observations and ssr is the sum of squared residuals. Since the variance is the same across all measurement equations, the measurement errors are pooled. We have $n = T(J-N)$, because at each point in time there are $J-N$ independent measurement errors, and $ssr = \sum_{t=1}^{T} ||Y_t - \hat{Y}_t||$. Furthermore, $\alpha_0 = \delta_0 = 0$ because the prior is taken to be completely diffuse.

E Model-selection samplers

This section describes the three model-selection samplers that can alternatively be used to draw (λ, γ). Sections 3 and 4 in the paper show the robustness of the model-selection results to the use of different sampling algorithms.

For a recent overview of methods available for joint model-parameter sampling, see Godsill (2001) and Sisson (2005). For a review of MCMC methods for variable selection, see O'Hara and Sillanpää (2009).

E.1 Gibbs Variable Selection

The first approach is based on Gibbs Variable Selection (GVS), which was developed by Dellaportas et al. (2002) and is a special case of the product-space sampling of Carlin and Chib (1995). In product-space sampling, the parameter space is the product-space of all models under consideration, which can become very large. For GVS, models are treated as nested, so that the product-space is simply the parameter space of the unrestricted model.

The parameter prior for GVS is

$$P(\lambda|\gamma) = P(\lambda_\gamma|\gamma)P(\lambda_{\setminus\gamma}|\lambda_\gamma, \gamma)$$

where $P(\lambda_\gamma|\gamma)$ is the prior for those elements of λ that are included, and $P(\lambda_{\setminus\gamma}|\lambda_\gamma, \gamma)$ denotes the prior for elements that are not currently included, the so-called "pseudo-prior" or "linking density" in the parlance of Carlin and Chib (1995). It is generally recommended to use a pseudo-prior similar to the posterior distribution from the unrestricted model. To this end, I use independent normal distributions with mean and variance equal to the conditional posterior moments of λ given the ML estimates of all other parameters.[2]

In each iteration of the GVS sampler, first λ_γ is drawn in a Gibbs step from its conditional posterior distribution. Due to prior conditional independence, this only depends on the prior for λ_γ, on γ, on the remaining parameters, and on the data. The conditional posterior distribution is given in Online Appendix C.2. For those elements of λ that are not included in the current model, $\lambda_{\setminus\gamma}$, the data is not informative and the values are drawn from the pseudo-prior.

The success probability for the Bernoulli conditional posterior of $\gamma_i^{(i)}$ is determined by

$$\begin{aligned}&\frac{P(\gamma_i^{(j)} = 1|\lambda^{(j)}, \theta_-^{(j-1)}, \gamma_{-i}^{(j)}, Y)}{P(\gamma_i^{(j)} = 0|\lambda^{(j)}, \theta_-^{(j-1)}, \gamma_{-i}^{(j)}, Y)} \\ &\quad = \frac{P(Y|\gamma_i^{(j)} = 1, \lambda^{(j)}, \theta_-^{(j-1)}, \gamma_{-i}^{(j)})}{P(Y|\gamma_i^{(j)} = 0, \lambda^{(j)}, \theta_-^{(j-1)}, \gamma_{-i}^{(j)})} \cdot \frac{P(\lambda_i^{(j)}|\gamma_i^{(j)} = 1)}{P(\lambda_i^{(j)}|\gamma_i^{(j)} = 0)} \cdot \frac{P(\gamma_i^{(j)} = 1, \gamma_{-i}^{(j)})}{P(\gamma_i^{(j)} = 0, \gamma_{-i}^{(j)})}.\end{aligned}$$

To calculate the first term, the ratio of likelihoods, only the $\mathbb{P}$-likelihoods need to be evaluated, since the parameters of the $\mathbb{Q}$-likelihood remain unchanged and this term cancels out. For the second term, the numerator is the density of the prior, i.e., a normal distribution with mean zero and variance v_i, and the denominator is the density of the pseudo-prior. The third term cancels out due to equal prior model probabilities.

E.2 Stochastic Search Variable Selection

Stochastic Search Variable Selection (George and McCulloch, 1993, SSVS) was the first MCMC sampling approach to variable selection. It was developed by George and McCulloch (1993) and has since been applied extensively, including for Bayesian VAR estimation (George et al.,

[2]This is a convenient alternative to the common practice of doing a pilot run for the unrestricted model, taking advantage of the closed-form availability of conditional posterior moments for λ.

2008). The idea here is that those parameters that are excluded from the model are taken to be from a prior that is a tight distribution around zero.

Formally, the parameter prior is specified as the following normal mixture: $P(\lambda_i|\gamma_i) = (1-\gamma_i)N(0,\tau_{0i}^2) + \gamma_i N(0,\tau_{1i}^2)$. This can be obtained from the multivariate normal prior $P(\lambda|\gamma) = N(0, D_\gamma R D_\gamma)$, where D_γ is a diagonal matrix with elements $(1-\gamma_i)\tau_{0i} + \gamma_i\tau_{1i}$, and R is a correlation matrix, here equal to the identity matrix due to prior conditional independence. Intuitively, for included elements the prior variance, τ_{1i}^2, is chosen large so that the posterior estimate is informed by the data, while for excluded elements the prior variance, τ_{0i}^2, is set to a small value, so that posterior estimates are close to zero. Following common practice, I choose these variances to be multiples of the variance of the least-squares estimates for the unrestricted model: $\tau_{0i} = c_0\hat{\sigma}_{\lambda_i}$ and $\tau_{1i} = c_1\hat{\sigma}_{\lambda_i}$, where c_0 and c_1 are tuning parameters. Specifically, I will set $c_1 = \sqrt{g}$, in line with the prior specification for the other samplers, and $c_0 = 1/c_1$. These choices leads to good convergence speed.

In iteration j, first λ is drawn conditional on the values for all other parameters from iteration $j-1$. All elements of λ are drawn simultaneously in a Gibbs step from the normal conditional posterior distribution, whose moments are given in Appendix C.2. Note that the prior covariance matrix is $D_\gamma R D_\gamma$ and no zeros are imposed on λ.

The elements of γ are drawn successively, in random order, from their conditional posterior distributions. Denote by $\gamma_i^{(j)}$ the element that is sampled, and by $\gamma_{-i}^{(j)}$ all remaining elements. The success probability of the Bernoulli conditional posterior is determined by:

$$\begin{aligned}\frac{P(\gamma_i^{(j)}=1|\lambda^{(j)},\theta_-^{(j-1)},\gamma_{-i}^{(j)})}{P(\gamma_i^{(j)}=0|\lambda^{(j)},\theta_-^{(j-1)},\gamma_{-i}^{(j)})} &= \frac{P(\lambda^{(j)}|\gamma_i^{(j)}=1,\gamma_{-i}^{(j)},\theta_-^{(j-1)})}{P(\lambda^{(j)}|\gamma_i^{(j)}=0,\gamma_{-i}^{(j)},\theta_-^{(j-1)})}\cdot\frac{P(\gamma_i^{(j)}=1)}{P(\gamma_i^{(j)}=0)} \\ &= \frac{(\tau_{1i})^{-1}\exp(-.5(\lambda_i^{(j)}/\tau_{1i})^2)}{(\tau_{0i})^{-1}\exp(-.5(\lambda_i^{(j)}/\tau_{0i})^2)},\end{aligned}$$

where the second line follows from the prior conditional independence of the elements of λ and the equal prior model probabilities. This ratio does not depend on the data because in the SSVS approach, γ affects the likelihood only through λ.

E.3 Reversible-jump Markov chain Monte Carlo

The Reversible-jump Markov chain Monte Carlo (RJMCMC) sampler developed by Green (1995) allows moves between parameter spaces of different dimensionality, and therefore can deal with the setting in which the prior of the excluded elements of λ is simply a point mass at zero. This approach is extremely flexible, allowing for different types of moves between

models. I adapt the local reversible-jump sampler of Dellaportas and Forster (1999), where only local moves are considered. Denote the current state of the chain by $(\lambda^{(j)}, \gamma^{(j)}, \theta_-^{(j)})$. First the decision is made between a "null move" (a within-model move) and a "jump move" (a between-model move). With 25% probability a null move is undertaken, in which γ remains unchanged and λ is updated using a Gibbs step. If a jump move is attempted, a proposal for (λ', γ') is constructed by *changing only one parameter*. The index of the element to be changed, i, is randomly chosen. If $\gamma_i^{(j)} = 0$, then the element is added to the model, otherwise it is deleted. When the move involves adding a parameter to the model, the proposal is $\lambda' = g(\lambda^{(j)}, u)$, where u is a scalar drawn from the proposal density $q_i(u)$, taken to be $N(\mu_i, \sigma_i^2)$, and $g(\cdot)$ is the identity transformation (such that $\lambda_i' = u$). The acceptance probability is

$$\begin{aligned}\alpha(\lambda^{(j)}, \gamma^{(j)}, \theta_-^{(j)}, \lambda', \gamma') &= \frac{P(Y|\lambda', \gamma', \theta_-^{(j)})}{P(Y|\lambda^{(j)}, \gamma^{(j)}, \theta_-^{(j)})} \frac{P(\lambda'|\gamma')}{P(\lambda^{(j)}|\gamma^{(j)})} \frac{P(\gamma')}{P(\gamma^{(j)})} \frac{1}{q_i(u)} \\ &= \frac{P(Y|\lambda', \gamma', \theta_-^{(j)})}{P(Y|\lambda^{(j)}, \gamma^{(j)}, \theta_-^{(j)})} \frac{v_i^{1/2} \exp(-u^2/v_i)}{\sigma_i \exp(-(u-\mu_i)^2/\sigma_i^2)}\end{aligned}$$

where the second line follows from prior conditional independence and equal prior model probability.[3] When an element is deleted from the model, we have $(\lambda', u') = g^{(-1)}(\lambda^{(j)})$, meaning that λ' now has a zero at the i-th position, and $\lambda_i^{(j)} = u'$. In this case,

$$\alpha(\lambda^{(j)}, \gamma^{(j)}, \theta_-^{(j)}, \lambda', \gamma') = \frac{P(Y|\lambda', \gamma', \theta_-^{(j)})}{P(Y|\lambda^{(j)}, \gamma^{(j)}, \theta_-^{(j)})} \frac{\sigma_i \exp(-(u'-\mu_i)^2/\sigma_i^2)}{v_i^{1/2} \exp(-u'^2/v_i)}.$$

The choice of the proposal distributions plays an important role for the efficiency of the sampler. Ideally, they should be close to the conditional posteriors (Goodsill, 2001, Sec. 2.3.2). Here, the parameter conditionals for each element of λ are in fact available in closed form, based on standard results.[4] I use the moments of these conditional posterior distributions, calculated anew in each iteration for the chosen element i, for μ_i and σ_i^2. The resulting sampler achieves good efficiency in that it jumps between models frequently.

[3] As for GVS, the ratio of the likelihoods only requires calculating the $\mathbb{P}$-likelihood.

[4] Consider the relevant equation in the system $X_t - E_{t-1}^{\mathbb{Q}} X_t = \lambda_0 + \lambda_1 X_{t-1} + \Sigma\varepsilon_t$. After subtracting all terms not involving the parameter of interest λ_i, the conditional posterior moments for λ_i follow from standard results for univariate Bayesian regression—see also Geweke (1996) and Kuo and Mallick (1998).

F Details and additional results for simulation study

The data used to estimate parameters for the DGP is described in Section 4. I estimate a two-factor DTSM on this data—the two factors correspond to level and slope of the yield curve. To construct the DGP, I first obtain the maximum likelihood estimates for the unrestricted model, and determine which elements of λ are significantly different from zero at the 5% level. Only one out of six parameters is significant. Then in a second step I obtain maximum likelihood estimates of the model which restricts the other five parameters to zero. The parameters estimated in this way are used in the DGP. The risk risk-price parameters are

$$\lambda_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \lambda_1 = \begin{pmatrix} 0 & -0.0375 \\ 0 & 0 \end{pmatrix},$$

and the remaining parameters are $k_\infty^{\mathbb{Q}} = 4.9 \cdot 10^{-5}$, $\phi^{\mathbb{Q}} = (0.9916, 0.9750)$, $\sigma_e = 2/120000$ (2 basis points), and $\Sigma = ((2.2 \cdot 10^{-4}, 0.6 \cdot 10^{-4})', (0, 1.2 \cdot 10^{-4})')$. The results are discussed in Section 3 of the paper.

In additional, unreported results, I have found that for longer sample sizes, the model-selection samplers become even more accurate in recovering the DGP specification when longer samples are simulated. This indicates that small samples make it more difficult to correctly identify risk price restrictions, consistent with the notion that small-sample bias plays a role for both parameter estimates and model selection.

While the DGP described above is chosen based on actual estimates in the data, with many risk-price parameters restricted to zero, it may be of some interest to also consider a DGP that leaves all risk-price parameters unrestricted. In this case, there is unfortunately no obvious way to choose the risk-price parameters. The ML estimates are close to zero for most risk-price parameters, and simulating from such a model, where elements of λ are non-zero but very small, would make it impossible for any model selection algorithm to identify all parameters as truly non-zero. The DGP risk-price parameters therefore have to be chosen in some other way. I do so by setting them to equal values in absolute magnitude, choosing this magnitude to be roughly economically reasonable, and taking into account the restriction that these values do not make the VAR explosive. This leads me to the following parameters:

$$\lambda_0 = 1200^{-1} \begin{pmatrix} .1 \\ .1 \end{pmatrix} \quad \lambda_1 = \begin{pmatrix} -.1 & -.1 \\ .1 & -.1 \end{pmatrix}.$$

The remaining DGP parameters are the ML estimates for the unrestricted model, which after rounding are the same as those reported above. The results for model selection in the simu-

lated yield data are shown in Table A.1. Bayesian model selection still does reasonably well, even though choosing restrictions from estimates of the unrestricted model is now somewhat more accurate. Specifically, choosing a preferred model based on 95%-CIs for the risk-price parameters from the unrestricted estimation leads to the correct, unrestricted model in 73 out of 100 simulated samples. Choosing the modal model based on the posterior model probabilities from SSVS, GVS, and RJMCMC leads to the correct model in only 46-48 out of 100 samples. The model prior plays a role in explaining this result. Since it puts equal probability on each parameter being unrestricted or restricted, it leads to a Binomial distribution over the number of unrestricted parameters with a mean of three (see Section 2.4). Hence even if the likelihood favors six unrestricted parameters, the posterior will tend to favor at least some restrictions due to the prior. It is not possible to specify a completely uninformative model prior in Bayesian variable selection. One could increase the prior probability of inclusion for the risk-price parameters, in order to favor less restrictive models (Chipman et al., 2001; Clyde and George, 2004) if there are *a priori* reasons to do so. However, in the context of term structure models there are no such reasons—on the contrary, one might want to instead impose more parsimony via the prior to make better use of no-arbitrage via additional risk-price restrictions.

In sum, the results of the simulation study show that my Bayesian model selection samplers do reasonably well for alternative DGP specifications. For a plausible DGP based on estimates from the data, this approach accurately infers the true model specification. It is possible to write down DGPs for which model choice based on posterior CIs from the estimation of an unrestricted model performs better than joint-model-parameter sampling, but the latter approach still performs satisfactory. Of course, only joint-model-parameter sampling can appropriately incorporate model uncertainty in this context.

G Excess bond returns

The continuously compounded return of holding an n-period bond for h periods, in excess of the risk-free rate, is

$$rx_{t,t+h}^{(n)} = -(n-h)y_{t+h}^{n-h} + ny_t^n - hy_t^h.$$

For fitted excess returns, calculated using model-implied yields, we have

$$\begin{aligned} \hat{rx}_{t,t+h}^{(n)} &= -(n-h)\hat{y}_{t+h}^{n-h} + n\hat{y}_t^n - h\hat{y}_t^h \\ &= \mathcal{A}_{n-h} - \mathcal{A}_n + \mathcal{A}_h + \mathcal{B}_{n-h}' X_{t+h} - (\mathcal{B}_n - \mathcal{B}_h)' X_t. \end{aligned}$$

Fitted excess returns are generally close to observed excess returns, because of the accurate cross-sectional yield fit of the models considered in this paper.

The time-t model-implied expected excess return is

$$\begin{aligned} E_t\hat{rx}_{t,t+h}^{(n)} &= \mathcal{A}_{n-h} - \mathcal{A}_n + \mathcal{A}_h + \mathcal{B}'_{n-h}E_tX_{t+h} - (\mathcal{B}_n - \mathcal{B}_h)'X_t \\ &= \mathcal{A}_{n-h} - \mathcal{A}_n + \mathcal{A}_h + \mathcal{B}'_{n-h}[(I_N - \Phi^h)E(X_t) + \Phi^h X_t] - (\mathcal{B}_n - \mathcal{B}_h)'X_t \\ &= \mathcal{A}_{n-h} - \mathcal{A}_n + \mathcal{A}_h + (\mathcal{B}'_{n-h} - \mathcal{B}'_n + \mathcal{B}'_h)E(X_t) + (\mathcal{B}'_{n-h}\Phi^h - \mathcal{B}'_n + \mathcal{B}'_h)(X_t - E(X_t)), \end{aligned}$$

and the surprise component of the excess return is

$$\hat{rx}_{t,t+h}^{(n)} - E_t\hat{rx}_{t,t+h}^{(n)} = \mathcal{B}'_{n-h}(X_{t+h} - E_tX_{t+h}) = \mathcal{B}'_{n-h}\sum_{i=1}^{h}\Phi^{h-i}\Sigma\varepsilon_{t+i},$$

see also equation (11) of Duffee (2011).

To gain some intuition about the model's implications for returns, consider the one-period excess return, which can be written as follows:

$$\hat{rx}_{t,t+1}^{(n)} = -\frac{1}{2}\mathcal{B}'_{n-1}\Sigma\Sigma'\mathcal{B}_{n-1} + \mathcal{B}'_{n-1}\Sigma\lambda_t + \mathcal{B}'_{n-1}\Sigma\varepsilon_{t+1}. \tag{4}$$

The first term, which corresponds to $\frac{1}{2}Var_t(\hat{rx}_{t,t+1}^{(n)})$, is due to convexity. The second term captures the actual risk compensation for level, slope, and curvature risk. This can be seen by rewriting it as $\lambda_t'Cov_t(\varepsilon_{t+1}, \hat{rx}_{t,t+1}^{(n)})$, the product of the prices of risk and the quantities of risk. In a Gaussian model, quantities of risk are constant, and time-variation in expected returns is due exclusively to movements in λ_t. The third term in equation (4) captures the surprise component of the excess return.

Consider now the predictability regression in Section 5.3, for which we want to derive the model-implied R^2. Because the regressors correspond to the risk factors in the DTSMs, the population R^2 is equal to the variance of model-implied expected excess returns divided by the variance of model-implied realized excess returns, i.e.,

$$R^2_{pop} = \frac{Var(E_t\hat{rx}_{t,t+h}^{(n)})}{Var(\hat{rx}_{t,t+h}^{(n)})}.$$

The variance of expected excess returns is given by

$$Var(E_t\hat{rx}_{t,t+h}^{(n)}) = (\mathcal{B}'_{n-h}\Phi^h - \mathcal{B}'_n + \mathcal{B}'_h)Var(X_t)(\mathcal{B}'_{n-h}\Phi^h - \mathcal{B}'_n + \mathcal{B}'_h)'.$$

The unconditional covariance matrix is calculated using $Var(X_t) = (I_{N^2} - \Phi \otimes \Phi)^{-1} vec(\Sigma\Sigma')$. The variance of realized excess returns is

$$Var(\hat{rx}^{(n)}_{t,t+h}) = Var(E_t \hat{rx}^{h}_{t,t+h}) + \mathcal{B}'_{n-h} \left(\sum_{i=1}^{h} \Phi^{h-i} \Sigma\Sigma' (\Phi^{h-i})' \right) \mathcal{B}_{n-h}.$$

Since yield loadings are very similar across models, differences in model-implied return predictability are due to differences in Φ, which in turn stem from differences in λ_1.

References

Carlin, Bradley P. and Siddartha Chib (1995) "Bayesian Model Choice via Markov Chain Monte Carlo Methods," *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 57, pp. 473–484.

Chib, S. and B. Ergashev (2009) "Analysis of Multifactor Affine Yield Curve Models," *Journal of the American Statistical Association*, Vol. 104, pp. 1324–1337.

Chib, Siddhartha and Edward Greenberg (1994) "Bayes Inference in Regression Models With ARMA(p,q) Errors," *Journal of Econometrics*, Vol. 64, pp. 183–206.

Chib, Siddhartha and Srikanth Ramamurthy (2010) "Tailored randomized block MCMC methods with application to DSGE models," *Journal of Econometrics*, Vol. 155, pp. 19–38.

Chipman, Hugh, Edward I. George, and Robert E. McCulloch (2001) "The Practical Implementation of Bayesian Model Selection," in P. Lahiri ed. *IMS Lecture Notes – Monograph Series*, Vol. 38: Institute of Mathematical Statistics, pp. 65–116.

Clyde, Merlise and Edward I. George (2004) "Model Uncertainty," *Statistical Science*, Vol. 19, pp. 81–94.

Dellaportas, Petros and Jonathan J. Forster (1999) "Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models," *Biometrika*, Vol. 86, pp. 615–633.

Dellaportas, Petros, Jonathan J. Forster, and Ioannis Ntzoufras (2002) "On Bayesian model and variable selection using MCMC," *Statistics and Computing*, Vol. 12, pp. 27–36.

Duffee, Gregory R. (2011) "Information In (and Not In) the Term Structure," *Review of Financial Studies*, Vol. 24, pp. 2895–2934.

Gamerman, Dani and Hedibert F. Lopes (2006) *Markov Chain Monte Carlo*: Chapman & Hall/CRC, 2nd edition.

George, Edward I. and Robert E. McCulloch (1993) "Variable Selection Via Gibbs Sampling," *Journal of the American Statistical Association*, Vol. 88, pp. 881–889.

George, Edward I., Dongchu Sun, and Shawn Ni (2008) "Bayesian stochastic search for VAR model restrictions," *Journal of Econometrics*, Vol. 142, pp. 553–580.

Geweke, John (1996) "Variable selection and model comparison in regression," in J. M. Bernando, J. O. Berger, A. P. Daviw, and A. F. M. Smith eds. *Bayesian Statistics 5*: Oxford University Press, pp. 609–620.

Godsill, Simon J. (2001) "On the Relationship Between Markov Chain Monte Carlo Methods for Model Uncertainty," *Journal of Computational and Graphical Statistics*, Vol. 10, pp. 230–248.

Goodsill, Simon J. (2001) "On the Relationship Between Markov Chain Monte Carlo Methods for Model Uncertainty," *Journal of Computational and Graphical Statistics*, Vol. 10, pp. 230–248.

Green, Peter J. (1995) "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination," *Biometrika*, Vol. 82, pp. 711–732.

Joslin, Scott, Marcel Priebsch, and Kenneth J. Singleton (2014) "Risk Premiums in Dynamic Term Structure Models with Unspanned Macro Risks," *Journal of Finance*, Vol. 69, pp. 1197–1233.

Kuo, Lynn and Bani Mallick (1998) "Variable selection for regression models," *Sankhya Ser. B*, Vol. 60, pp. 65–81.

Lütkepohl, Helmut (2006) *New introduction to multiple time series analysis*: Springer Verlag.

O'Hara, Robert B. and Mikko J. Sillanpää (2009) "A Review of Bayesian Variable Selection Methods: What, How and Which," *Bayesian Analysis*, Vol. 4, pp. 85–118.

Sisson, Scott A. (2005) "Transdimensional Markov Chains: A Decade of Progress and Future Perspectives," *Journal of the American Statistical Association*, Vol. 100, pp. 1077–1090.

Zellner, Arnold (1962) "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias," *Journal of the American Statistical Association*, Vol. 57, pp. 348–368.

Zellner, Arnold and Tomohiro Ando (2010) "A direct Monte Carlo approach for Bayesian analysis of the seemingly unrelated regression model," *Journal of Econometrics*, Vol. 159, pp. 33–45.

Table A.1: Simulation study for alternative unrestricted DGP: risk-price restrictions

	Element of γ						Freq. of
	(1)	(2)	(3)	(4)	(5)	(6)	corr. model
DGP	1	1	1	1	1	1	
MCMC	0.99	1.00	0.92	1.00	0.81	1.00	73%
SSVS	0.86	0.99	0.75	0.96	0.69	0.96	46%
GVS	0.86	0.99	0.75	0.95	0.69	0.96	47%
RJMCMC	0.85	0.99	0.72	0.94	0.71	0.95	48%

Risk-price specification of the data-generating process (DGP), and estimation results in simulated data. For MCMC (estimation of unrestricted model), the fraction of samples in which the 95%-credibility interval for the corresponding element of λ did not straddle zero. For model-selection samplers, average posterior means for γ across simulations. Last column shows the percentage of samples in which the correct model was chosen—for MCMC model choice is based on 95%-credibility intervals and for the model-selection samplers on posterior model probabilities.