A Computationally Efficient Mixture Innovation Model for Time-Varying Parameter Regressions

Zhongfang He*

First Version: September 12, 2021

Abstract

The mixture innovation (MI) model places spike-and-slab mixture distributions on the innovations of time-varying regression coefficients and permits flexible time variations while allowing for dynamic shrinkage. Despite its appeal, the standard algorithm of Gerlach et al. (2000) requires $\mathcal{O}(2^K n)$ operations to draw the mixture indicators in the MI model for n observations with K regressors and is computationally prohibitive when K grows. As an alternative, this paper proposes a new specification of the MI model in which the 0/1 mixture indicators in the original MI model are approximated by a logistic function of latent continuous variables. Through Metropolis-Hastings steps, the latent variables and hence the approximated mixture indicators can be sampled at an $\mathcal{O}(n)$ cost, thus offering large improvement in computation efficiency while keeping the benefits of the MI model. An efficient MCMC algorithm is developed to estimate the new model. I show in a simulation study that the new model can achieve the same level of estimation accuracy as the original MI model but at a much smaller computation cost. The new model is further tested in two empirical applications with a relatively large number of regressors for which the original MI model is practically infeasible.

Keywords: TVP, Dynamic Shrinkage, Spike-and-Slab Prior, Mixture Innovation
JEL Codes: C11, C22, E37, G17

*Email: hezhongfang2004@yahoo.com. Royal Bank of Canada, 155 Wellington St W, Toronto, ON, Canada, M5V 3H6. The views in this paper are solely the author's responsibility and are not related to the company the author works in.

1 Introduction

Many research efforts in studies of economic time series have been devoted to allow time variations in the parameters of regression models for greater flexibility and improved forecasts. Examples include Hamilton (1989), Chib (1998), Primiceri (2005), Koop and Potter (2007), Maheu and Gordon (2008), Fruhwirth-Schnatter and Wagner (2010) and many others. An important theme of this literature is how to encourage model parsimony while allowing model flexibility. The Bayesian approach to address this problem is by imposing shrinkage priors for the time varying parameters. To set the stage, consider the time-varying parameter (TVP) regression $y_t = x_t'\beta_t + \epsilon_t$ with $\beta_t = \beta_{t-1} + \eta_t$, where y_t is a scalar dependent variable, x_t is a K-dimensional vector of regressors, β_t is the corresponding time-varying coefficients with a starting value β_0, ϵ_t and $\eta_t = [\eta_{1,t} \ ... \ \eta_{K,t}]'$ are the innovations of the dependent variable and the regression coefficients respectively for a time series sample $t = 1, 2, ..., n$. Substituting $\beta_t = \beta_0 + \eta_1 + ... + \eta_t$ in the TVP regression leads to an expanded linear regression:

$$y_t = x_t'\beta_0 + x_{\eta,t}'\eta + \epsilon_t \tag{1}$$

where $x_{\eta,t} = [\mathbf{1}_t' \otimes x_t' \ \ \mathbf{0}_{(n-t)K}']'$ is a nK dimensional vector, $\mathbf{1}_t$ is a t-by-1 vector of ones, $\mathbf{0}_{(n-t)K}$ is a $(n-t)K$-by-1 vector of zeros, the symbol $\otimes$ denotes the Kronecker product and $\eta = [\eta_1' \ ... \ \eta_n']'$ is a nK dimensional vector stacking the coefficient innovations η_t for $t = 1, ..., n$. In particular, Equation (1) shows that the coefficient innovations η can be treated as the static coefficients of a linear regression. Hence one can connect with the rich literature of Bayesian shrinkage regressions to develop shrinkage priors for the coefficient innovations η of a TVP regression to discourage unnecessary parameter variations.

A prominent regression shrinkage method in the Bayesian literature is the spike-and-slab approach (Mitchell and Beauchamp (1988), George and McCulloch (1993), Kuo and Mallick (1998), Ishwaran and Rao (2005)) that is often considered as a "methodological ideal" of

shrinkage estimation (Carvalho et al. (2009)). The spike-and-slab approach places a prior of a two-component mixture distribution for each regression coefficient with a point mass at zero (or a very narrow distribution around zero) and a relatively diffuse distribution. The resulting posterior distribution effectively performs Bayesian model averaging over regressor combinations (Hoeting et al. (1999)). Placing a spike-and-slab prior for the static coefficient η in the linear regression of Equation (1) leads to a version of the mixture innovation (MI) model of Giordani and Kohn (2008) that sets $\eta_{j,t} \sim N(0, d_{j,t}v_j^2)$ for each element in η where $d_{j,t} \in \{0, 1\}$ is a Bernoulli variable of the mixture indicator, v_j^2 is the variance of the slab part over $j = 1, ..., K$ and $t = 1, ..., n$. The mixture indicator $d_{j,t}$ switches on and off the innovation $\eta_{j,t}$ by moving between values of one and zero to determine whether there is a random parameter shift or no parameter change locally at each time t and thus permits very flexible time variation patterns in the regression coefficients. For example, both infrequent change points and episodical combinations of constant and time-varying coefficients can be readily accommodated in the MI model. By letting the mixture indicator $d_{j,t} = 0$, the MI model is able to shut off unnecessary parameter shifts adaptively and enforce dynamic shrinkage.

Despite its appeal, estimating the MI model is challenging. A standard Gibbs sampler cycling through the coefficient innovation η and the mixture indicator $d = \{d_t\}_{t=1}^n$ with $d_t = [d_{1,t} \ ... \ d_{K,t}]'$ can break down completely due to the high correlation between η and d (Gerlach et al. (2000)). A solution to this sampling problem is by adopting the algorithm of Gerlach et al. (2000) (*GCK* hereafter) that draws the mixture indicator d by marginalizing over η in efficient $\mathcal{O}(n)$ operations. However, the resulting sampler suffers from the curse of dimensionality due to the fact that sampling a K-dimensional 0/1 variable d_t needs to evaluate its posterior distribution over all of its 2^K possible values. The total computation cost of sampling the mixture indicator d is $\mathcal{O}(2^K n)$ and can be prohibitive when K is even modestly large. In practical applications of the MI model, researchers often have to impose

ad hoc restrictions on the possible scenarios the mixture indicator can take in order to make the model estimation feasible (Chan et al. (2012)).

Giordani and Kohn (2008) proposes an adaptive Metropolis-Hastings (MH) algorithm to speed up the sampling of the mixture indicator in MI models. The GCK algorithm is first run for an initial training period to form an approximation of the posterior distribution of the mixture indicators. Subsequently the approximation is used as the proposal distribution in MH steps to sample the mixture indicators and is successively updated. To apply this algorithm, the researcher would need to determine the appropriate length of the initial training period carefully on a case-by-case basis. Moreover, when K is large, running the GCK algorithm even in the initial training period could be computationally expensive.

The root of the sampling difficulty in the MI model lies in the discrete nature of the mixture indicators. If they were continuous variables, one does not have to evaluate the model likelihood function under all possible values of the mixture indicators. Instead simple MH steps can be used to explore the parameter space and sample these "continuous" mixture indicators at a much lower computation cost. This observation motivates a new specification of the MI model where each component in the mixture indicator is an indicator function of a latent continuous variable. Hence the problem of sampling a vector of discrete mixture indicators in the original MI model is transformed into one of sampling a vector of continuous variables, which can be cheaply achieved by an MH step.

Specifically a "soft" version of the indicator function, namely the logistic function, is used to link the mixture indicators to auxiliary latent Gaussian variables. When a latent variable takes a large value at time t, its logistic function approximately equals one and hence turns on the innovation of the corresponding time-varying regression coefficient to capture possible parameter shift at time t. Similarly when the latent variable goes towards $-\infty$ at time t, its logistic function becomes close to zero and effectively switches off parameter change at time t. Compared to the original MI model, the latent variables

fully determine the mixture indicators in the new specification. The distribution of the resulting mixture indicator in the new MI model is analyzed and is found to be close to an Bernoulli one with probability mass concentrated near the points of 0 and 1.

An efficient MCMC scheme is developed to estimate the new MI model. Sampling the auxiliary latent variables in the new model is by a single-move Metropolis-within-Gibbs step cycling through $t = 1, 2, ..., n$. The model likelihood function that integrates out the coefficient innovation η is computed based on the GCK algorithm. As a result, sampling the latent variables and hence the mixture indicators at each time t requires evaluating the integrated model likelihood function only twice during an MH step, instead of 2^K times in the original MI model. The overall cost of sampling the mixture indicator d is $\mathcal{O}(n)$, which is a drastic reduction from the $\mathcal{O}(2^K n)$ cost in the original MI model.

In a simulation study with a relatively small number of regressors, the new model is found to produce estimates of the regression coefficients with the same level of accuracy as the original MI model but reduces the running time significantly. The new model is further tested in two empirical exercises of predicting equity premium and inflation rate. The improvement in computation efficiency is substantial. For example, in the equity premium prediction exercise that contains 14 regressors, producing 1,000 posterior draws from the new model takes around 120 seconds while simulating the same number of draws by the original MI model would require over 66 hours. In the inflation rate prediction exercise with 22 regressors, the new model takes around 140 seconds to simulate 1,000 draws. In contrast, the original MI model would require over 16 hours to produce a single draw in this case. In out-of-sample forecasts, the new model is found to outperform a version of the original MI model that imposes restrictions on the mixture indicator for computational feasibility as well as the dynamic horseshoe model of Kowal et al. (2019) that applies dynamic shrinkage to TVP models through an extension of the popular horseshoe prior of Carvalho et al. (2010).

The remainder of the paper is structured as follows. Section 2 describes the new MI model. The estimation algorithm is provided in Section 3. Section 4 and 5 present the simulation study and the empirical applications respectively. Section 6 concludes. Additional details are provided in appendices.

2 The Model

Following the discussions from Section 1, the TVP model under study is formally:

$$\begin{aligned} y_t &= x_t'\beta_t + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma_t^2), \\ \beta_{j,t} &= \beta_{j,t-1} + \eta_{j,t}, \quad \eta_{j,t} \sim N(0, d_{j,t} v_j^2), \\ \beta_{j,0} &\sim N(0, v_{j,0}^2), \ j = 1, ..., K \end{aligned} \tag{2}$$

The innovation ϵ_t of the dependent variable follows a normal distribution that allows time-varying conditional variance σ_t^2 by a stochastic volatility (SV) model:

$$h_t \equiv \log(\sigma_t^2) = (1 - \rho_h)\mu_h + \rho_h \, h_{t-1} + \epsilon_{h,t}, \ \ \epsilon_{h,t} \sim N(0, \sigma_h^2), \tag{3}$$

where $h_1 \sim N(\mu_h, \sigma_h^2/(1-\rho_h^2))$. The homoskedastic case can be accommodated by specifying $\sigma_t^2 = \sigma^2$. The initial value $\beta_{j,0}$ of each regression coefficient follows a zero-mean normal distribution with the variance $v_{j,0}^2$ for $j = 1, ..., K$.

The MI model of Giordani and Kohn (2008) can be obtained by specifying $d_{j,t}$ of Equation (2) as a Bernoulli variable which is equivalent to a spike-and-slab prior for each coefficient innovation $\eta_{j,t}$ for $j = 1, ..., K$ and $t = 1, ..., n$. To avoid the exponentially growing computation cost with respect to the number of regressors K, a new MI model is proposed

that specifies a hierarchical structure for $d_{j,t}$:

$$
\begin{aligned}
d_{j,t} &= \frac{1}{1+\exp(-z_{j,t})}, \\
z_{j,t} &= (1-\rho_j)\mu_j + \rho_j z_{j,t-1} + \xi_{j,t}, \quad \xi_{j,t} \sim N(0, a_j^2), \\
z_{j,1} &\sim N(\mu_j, a_j^2/(1-\rho_j^2)),\ j = 1, ..., K
\end{aligned} \tag{4}
$$

where $z_{j,t}$ is a latent variable following a first-order autoregressive (AR) process with the long-run mean μ_j, AR coefficient ρ_j and conditional variance a_j^2, which intends to capture possible serial persistence in the mixture indicator. The variable $d_{j,t}$ in Equation (4) is a logistic function of the latent variable $z_{j,t}$.[1] It is well known that when the absolute value of the latent variable $z_{j,t}$ is large, the variable $d_{j,t}$ takes value close to zero and one and effectively plays the same role as the 0/1 mixture indicator in the original MI model, thus allowing a close approximation to the original MI model to obtain its theoretical benefits. To keep the terminology simple, the variable $d_{j,t}$ in Equation (4) is still referred to as a mixture indicator even though it is now continuous between 0 and 1. The model of Equations (2) and (4) is referred to as a logistic MI (*LMI* hereafter) model. As will be shown in Section 3, the mixture indicators d in the LMI model can be sampled at an $\mathcal{O}(n)$ computational cost, which is a significant reduction from the $\mathcal{O}(2^K n)$ cost of the original MI model when the number of regressors K grows.

A few remarks are in order. Instead of the logistic function, a hard indicator function could be used to set elements in d_t to zero or one when corresponding elements in the latent variable $z_t = [z_{1,t}\ ...\ z_{K,t}]'$ cross some threshold and thus exactly match the mixture indicator in the original MI model. The downside is that only the sign of z_t will enter the model likelihood function, which can be seen by integrating out the regression coefficient β_t by Kalman filter. As a result, the model likelihood function is flat with respect to z_t

[1]Note that there is no intercept or slope for $z_{j,t}$ as opposed to the general form of a logistic function in order to identify the location and scale of $z_{j,t}$.

with the only change at the point of the threshold. In my experiments, the posterior draws of z_t as well as other model parameters converge rather poorly when the hard indicator function is used.

In principal, the CDF of any probability distribution with support in the real line can be used to link the mixture indicator d_t and the latent variable z_t. An obvious alternative to the logistic function would be the Gaussian CDF. Also I focus on the case of univariate time series in this paper. Extension to the multivariate case is possible by adopting the time-varying Cholesky decomposition introduced in Carriero et al. (2019) and Lopes et al. (2021) that transforms a multivariate system into a set of independent univariate regressions.[2]

2.1 Prior of Parameters

Gamma priors[3] are specified for the variance parameters: $v_j^2 \sim G(0.5, 2\tau_{v,j})$, $a_j^2 \sim G(0.5, 2\tau_{a,j})$, that are equivalent to normal priors for the signed square root of the variance parameters $v_j = \pm\sqrt{v_j^2} \sim N(0, \tau_{v,j})$ and $a_j = \pm\sqrt{a_j^2} \sim N(0, \tau_{a,j})$ for $j = 1, ..., K$. As the mixture indicator is already a shrinkage device, no hierarchical structure is placed on the hyper-parameters $\tau_{v,j}$ and $\tau_{a,j}$ to keep the model simple. $\tau_{v,j}$ and $\tau_{a,j}$ are set at 10 to form weakly diffused priors for v_j and a_j for $j = 1, ..., K$.

The initial value $\beta_0 = [\beta_{1,0} \ ... \ \beta_{K,0}]'$ plays the role of fixed regression coefficients as seen in Equation (1). The horseshoe prior of Carvalho et al. (2010) is adopted to encourage insignificant elements in β_0 shrunken towards zero. Specifically the prior variance of $\beta_{j,0}$ is $v_{j,0}^2 = \tau_0\tau_j$ with $\tau_0 \sim IB(0.5, 0.5)$ and $\tau_j \sim IB(0.5, 0.5)$ for $j = 1, ..., K$, where IB denotes

[2]A subtle issue of the time-varying Cholesky decomposition is the "prior ordering" problem described in Carriero et al. (2019) that is related to the sensitivity of the triangular decomposition to the ordering of the dependent variables.

[3]The gamma distribution $G(\alpha, \beta)$ for a generic variable x has the density $\frac{1}{\Gamma(\alpha)\beta^\alpha}x^{\alpha-1}\exp(-\frac{x}{\beta})$.

the inverted beta distribution[4].

For the SV specification of the variance σ_t^2 of the dependent variable (Equation (3)), the priors are $\mu_h \sim N(0, 10)$, $\rho_h \sim N(0.95, 0.04)I_{\{-1<\rho_h<1\}}$, $\sigma_h^2|s_h \sim G(0.5, 2s_h)$ and $s_h \sim IB(0.5, 0.5)$. In the homoskedastic case $\sigma_t^2 = \sigma^2$, the Jeffery's prior $\sigma^2 \propto \frac{1}{\sigma^2}$ is used.

For the other parameters in the process of the latent variable $z_{j,t}$, a scale-mixing normal prior is placed for the long-run mean μ_j that is relatively tight but allows some degree of flexibility through the scale mixture: $\mu_j|\psi_j \sim N(0, \psi_j)$, $\psi_j \sim IG(10, 1)$, where IG denotes the inverse gamma distribution and $j = 1, ..., K$. The AR coefficient has a relatively diffuse prior $\rho_j \sim N(0, 1)I_{\{-1<\rho_j<1\}}$ over $j = 1, ..., K$ to reflect the *a priori* uncertainty over its possible value.

2.2 Distribution of the Mixture Indicator d_t

Without loss of generality, the shrinkage property of the mixture indicator d_t is examined for the case where the AR coefficients $\rho_j = 0$ for $j = 1, ..., K$, which is equivalent to the stationary distribution of the latent variable z_t. The general cases can be analyzed in the same way by conditioning d_t on the lag z_{t-1} and redefining the conditional mean of z_t as a linear function of z_{t-1}. For ease of exposition, the discussion here focuses on the single-regressor case where d_t and z_t are scalars.

Given the hierarchical structure of d_t and the distributional assumptions placed for z_t and its conditional mean μ and conditional variance a^2, the fully marginalized distribution of d_t is not available in closed form. Instead I examine the distribution of d_t that is conditional on μ but marginalizes over a^2. By utilizing the prior distribution $a^2 \sim G(0.5, 2\tau_a)$

[4]The density of an inverted beta distribution $IB(a, b)$ is $p(x) = \frac{x^{a-1}(1+x)^{-a-b}}{B(a,b)} I\{x > 0\}$ where $B(\cdot, \cdot)$ is the beta function and a and b are positive real numbers. If $x \sim IB(0.5, 0.5)$, then $\sqrt{x} \sim C^+(0, 1)$ and vice versa, where $C^+(0, 1)$ is a standard half-Cauchy distribution with the density $p(z) = \frac{2}{\pi(1+z^2)} I\{z > 0\}$.

and the density transformation from z_t to d_t, one can derive:

$$p(d_t|\mu,\tau_a) = \frac{1}{\pi\sqrt{\tau_a}d_t(1-d_t)}\mathcal{K}_0\left(\sqrt{\frac{(\log(d_t)-\log(1-d_t)-\mu)^2}{\tau_a}}\right)$$

where $\mathcal{K}_0(\cdot)$ denotes the modified Bessel function of the second kind with the order 0. It is illuminating to investigate this partially marginalized distribution of d_t over different values of μ and the hyper-parameter τ_a of a^2 in order to understand the shrinkage property of d_t and to motive the choice of the hyper-parameter τ_a.

Figure 1 shows the density $p(d|\mu,\tau_a)$ over the grid $\mu \in \{-0.5, 0, 0.5\}$ and $\tau_a \in \{2, 4, 10\}$. It can be seen that the hyper-parameter τ_a controls the overall shrinkage ability of d_t where a larger τ_a pushes more probability mass of d_t towards the two ends at 0 and 1. When τ_a is sufficiently large (e.g. 10), essentially all probability mass of d_t is located at the two points 0 and 1, hence behaving like a Bernoulli variable. On the other hand, the mean parameter μ affects the degree of asymmetry in the conditional density of d_t. When $\mu = 0$, the conditional density of d_t is symmetric around the central point 0.5. As μ moves towards the negative domain, more probability mass of d_t is allocated to be near the end of 0 and hence exerts stronger shrinkage. Conversely d_t becomes more concentrated near the end of 1 when μ moves towards the positive domain.

2.3 Closely Related Models

Many previous studies have explored the use of a hard indicator function or a smooth transition function such as the logistic one to model regime changes in regression parameters. Examples in this strand of literature include the threshold regression model of Tong (1990) and the smooth transition regression model of Terasvirta (1998) that employ exogenous variables or lagged dependent variable to trigger regime changes. A recent example is Chang et al. (2017) that uses a latent variable as the driver of regime changes. This paper differs from Chang et al. (2017) in that the indicator function of the LMI model operates on

the innovation of regression coefficients rather than directly on the level of the coefficients and allows a much wider range of possible time variation patterns in the coefficients.

Nakajima and West (2013) applies the hard indicator function as a shrinkage device to push "small" regression coefficients to zero locally at each time t and thus allows episodical combinations of smooth variations and constant of zero in regression coefficients. In contrast, the LMI model can accommodate episodical constant regression coefficients at levels other than zero and provides greater flexibility[5].

Many existing studies of shrinkage for TVP models focus on the the case of homoskedastic time varying coefficients. Examples include Fruhwirth-Schnatter and Wagner (2010), Belmonte et al. (2014), Bitto and Fruhwirth-Schnatter (2019), Cadonna et al. (2020) etc.[6] Using the notation from this paper, the coefficient innovation $\eta_{j,t}$ would follow $N(0, v_j^2)$ with a constant conditional variance. Through shrinkage, each regression coefficient $\beta_{j,t}$ can be either a constant (i.e. $v_j = 0$) or continuously time varying (i.e. $v_j > 0$). However other time variations such as infrequent change points or episodical combinations of constant and time-varying patterns can not be accommodated as in the MI and LMI models. Viewed from the static representation of Equation (1), the assumption of homoskedastic $\eta_{j,t}$ amounts to a highly structured prior for η where each set $\{\eta_{j,t}\}_{t=1}^n$ is controlled by a single hyper-parameter v_j^2, while the MI and LMI models allow dynamic shrinkage of $\{\eta_{j,t}\}_{t=1}^n$ at each time point t for $j = 1, ..., K$.

In the context of dynamic shrinkage for TVP models, there are recent studies that employ absolutely continuous shrinkage priors for the innovations of time-varying parameters (Hauzenberger et al. (2020), Huber and Pfarrhofer (2021)) as a computationally attractive alternative to the spike-and-slab approach in the MI model. Among these studies an

[5]Another example is Huber et al. (2019) that applies a thresholding approximation in MCMC draws of the original MI model to speed up computation but suffers from convergence issues (Dufays et al. (2021)).

[6]Also see Chan et al. (2020) for an example of specifying a full and time-invariant conditional covariance matrix for time varying coefficients.

influential approach is the dynamic horseshoe model of Kowal et al. (2019) that can be essentially obtained by replacing the logistic function in the LMI model of Equation (4) by an exponential function and additionally imposing a z distribution for the innovation $\xi_{j,t}$ of each latent variable $z_{j,t}$. In this paper, the forecast performance of the dynamic horseshoe model and the proposed LMI model will be compared in the empirical exercises.

Another notable approach for exerting dynamic shrinkage for TVP models is the dynamic normal-gamma model of Kalli and Griffin (2014) where each time-varying coefficient is the product of a "normalized" time-varying coefficient and a scaling factor. While the normalized coefficient follows an AR process with constant conditional variances as in conventional TVP models, the scaling factor is also time varying and is able to push the overall coefficient close to zero locally at each time t.[7] Compared to the dynamic normal-gamma model, the LMI model allows the time-varying coefficients to be locally constant at both zero and non-zero values and hence is more flexible. Additionally, the LMI model is computationally efficient and avoids the computational inefficiency of the dynamic normal-gamma model documented in Kowal et al. (2019).

3 Estimation

3.1 Gibbs Sampler for the LMI Specification

For estimation efficiency, Equation (4) is reparametrized as:

$$
\begin{aligned}
d_{j,t} &= \frac{1}{1+\exp(-\mu_j - a_j z^*_{j,t})}, \\
z^*_{j,t} &= \rho_j z^*_{j,t-1} + \xi^*_{j,t}, \quad \xi^*_{j,t} \sim N(0,1), \\
z^*_{j,1} &\sim N\left(0, 1/(1-\rho_j^2)\right)
\end{aligned} \tag{5}
$$

[7] The original model of Kalli and Griffin (2014) specifies a gamma autoregressive process for the scaling factor. Alternatively Uribe and Lopes (2020) assumes a spike-and-slab distribution for the scaling factor.

where $z^*_{j,t} = \frac{z_{j,t}-\mu_j}{a_j}$ is the normalized latent variable for $j = 1, ..., K$. The benefit of the reparameterization is that the parameters μ_j and a_j are moved into the equation of the mixture indicator $d_{j,t}$ that directly connects with the dependent variable y_t after integrating out β_t by Kalman filter and hence can be sampled more efficiently.

Let $y = \{y_t\}_{t=1}^n$, $x = \{x_t\}_{t=1}^n$, $\beta = \{\beta_t\}_{t=1}^n$, $z^* = \{z^*_t\}_{t=1}^n$ with $z^*_t = [z^*_{1,t} \ ... \ z_{K,t}]'$, $\rho = [\rho_1 \ ... \ \rho_K]'$, $\mu = [\mu_1 \ ... \ \mu_K]'$, $a = [a_1 \ ... \ a_K]'$, $v = [v_1 \ ... \ v_K]'$, $\theta = \{\rho, \mu, a, v, \beta_0\}$, $\theta_0 = \{\tau_0, \tau, \psi\}$ with $\tau = [\tau_1 \ ... \ \tau_K]'$ and $\psi = [\psi_1 \ ... \ \psi_K]'$, θ_σ include σ^2 in the homoskedastic case or $\{\sigma^2_t\}_{t=1}^n$ and associated other parameters in the SV model, and $\Theta = \{\theta, \theta_0, \theta_\sigma\}$. The target is to sample from the posterior distribution $p(\beta, z^*, \Theta|y, x)$. The model parameters are divided into 3 blocks θ_0, θ_σ and $\{\beta, z^*, \theta\}$ to apply a Gibbs sampler.

The block θ_0 contains the estimable hyper-parameters of the priors for β_0 and μ. Conditional on β_0, the posterior of its hyper-parameters τ_0 and τ is inverse gamma distributions based on the hierarchical inverse gamma representation of the inverted beta distribution in Makalic and Schmidt (2016). The details of sampling τ_0 and τ are provided in Appendix A. Conditional on μ, the posterior of its hyper-parameters ψ_j can be easily derived as $IG(10.5, 1 + \frac{\mu_j^2}{2})$ for $j = 1, ..., K$.

In the case of the SV model, the sampler of Kastner and Fruhwirth-Schnatter (2014) can be readily applied to sample θ_σ. The details are provided in Appendix B. In the homoskedastic case $\sigma^2_t = \sigma^2$ with the prior $\sigma^2 \propto \frac{1}{\sigma^2}$, the posterior is $\sigma^2|y, x, \beta \sim \text{IG}\left(\frac{n}{2}, \frac{1}{2}\sum_{t=1}^n \epsilon_t^2\right)$ where $\epsilon_t = y_t - x_t'\beta_t$.

For the block $\{\beta, z^*, \theta\}$, this paper follows Gerlach et al. (2000) and samples z^* by integrating out β to avoid the sampling inefficiency due to the high dependence between β and z^*. Specifically, the posterior is decomposed as:

$$p(\beta, z^*, \theta|y, x, \theta_0, \theta_\sigma) = p(z^*, \theta|y, x, \theta_0, \theta_\sigma)p(\beta|y, x, z^*, \Theta)$$

The part $p(\beta|y, x, z^*, \Theta)$ can be sampled as the latent states in a linear Gaussian state space

system by the simulation smoother of Durbin and Koopman (2002)[8]. For the remaining piece $p(z^*, \theta|y, x, \theta_0, \theta_\sigma)$ that integrates out β, a nested Metropolis-within-Gibbs sampler is applied to iterate over the following sub-blocks:

1. Block z^*: A single-move Gibbs sampler is applied to sample from:

$$p(z_t^*|y, x, \Theta, z_{-t}^*) \propto p(z_t^*|z_{-t}^*, \rho)p(y|x, z^*, \Theta)$$

The prior part $p(z_t^*|z_{-t}^*, \rho)$ is a normal distribution thanks to the AR specification of z_t^*. The likelihood part can be written as $p(y|x, z^*, \Theta) = \prod_{t=0}^{n-1} p(y_{t+1}|y^t, x, z^*, \Theta)$ where $y^t = \{y_1, ..., y_t\}$ and y^0 is an empty set. Each component $p(y_{t+1}|y^t, x, z^*, \Theta)$ in the likelihood function can be derived as a normal distribution by applying a Kalman filter to integrate out β. However, the computation cost would be too high to run the Kalman filter for z_t^* at each time point $t = 1, ..., n$.

The computation efficiency can be greatly improved by noticing that only the components of the likelihood function relevant to z_t^* need to be computed at each time step t. One can apply the factorization $p(y|x, z^*, \Theta) \propto p(y_t, ..., y_n|y^{t-1}, x, z^*, \Theta)$. The right-hand side of the factorization can be computed efficiently by the GCK algorithm. The details are provided in Appendix C.

It is worth noting that, to sample z_t^* and hence the mixture indicator d_t in the LMI model, the posterior $p(z_t^*|y, x, \Theta, z_{-t}^*)$ only needs to be evaluated twice via the GCK algorithm in an MH step. In contrast, directly sampling the 0/1 mixture indicator d_t in the original MI model requires to evaluate the posterior of d_t by the GCK algorithm over all its 2^K possible scenarios. As a result, the overall computation cost of drawing $d = \{d_t\}_{t=1}^n$ is $\mathcal{O}(2^K n)$ in the original MI model but is reduced to be $\mathcal{O}(n)$ in the LMI model.

[8]Alternative approaches to simulate the latent states from a linear Gaussian state space system include Rue (2001) and McCausland et al. (2011) etc.

2. Block ρ: The posterior is $p(\rho|y, x, z^*, \Theta_{-\rho}) \propto p(\rho)p(z^*|\rho)$ where $\Theta_{-\rho}$ removes ρ from Θ. Given the truncated normal priors for each ρ_j, the kernel of the resulting posterior is $(1-\rho_j^2)^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\left(\rho_j^2(1-(z_{j,1}^*)^2)+\sum_{t=1}^{n-1}(z_{j,t+1}^*-\rho z_{j,t}^*)^2\right)\right) I_{\{-1<\rho_j<1\}}$ for $j = 1$, ..., K. Independent MH steps are used in the sampling. The proposal for each ρ_j arises from an auxiliary regression that ignores the stationarity constraint and the distribution of the initial value $z_{j,1}^*$ and is a normal distribution with the mean $\sum_{t=1}^{n-1} z_{j,t}^* z_{j,t+1}^*/(1+\sum_{t=1}^{n-1}(z_{j,t}^*)^2)$ and the variance $1/(1+\sum_{t=1}^{n-1}(z_{j,t}^*)^2)$ for $j = 1$, ..., K.

3. Blocks μ, a, v, β_0: : For the parameter block μ, the posterior is $p(\mu|y, x, z^*, \Theta_{-\mu}) \propto p(\mu|\psi)p(y|x, z^*, \Theta)$ where $\Theta_{-\mu}$ removes μ from Θ. The prior $p(\mu|\psi)$ is a normal distribution and is described in Section 2.1. As in the previous discussion of drawing z_t^*, the likelihood $p(y|x, z^*, \Theta)$ is a product of normal distributions and can be computed by applying a Kalman filter to integrate out β. Unlike in the case of z_t^*, calculating the likelihood $p(y|x, z^*, \Theta)$ for drawing μ does not need to be repeated over $t = 1$, ..., n and hence is computationally affordable. Details of computing the likelihood $p(y|x, z^*, \Theta)$ are provided in Appendix D.

   Drawing the other parameter blocks a, v, β_0 is similar to the block μ. The priors for a, v, β_0 are all normal distributions and are described in Section 2.1, while computing the likelihood $p(y|x, z^*, \Theta)$ is the same as in the step for μ.

To avoid manual tuning, the adaptive optimal scaling method of Garthwaite et al. (2016) is adopted in the MH steps. Take z_t^* for example. The proposal for its $i+1^{\text{th}}$ draw is a random walk $z_t^*(i+1) \sim N(z_t^*(i), w_i^2 A)$ where A equals I_K when $i \leq i^*$ and the sample covariance matrix $\frac{1}{i}\sum_{j=1}^{i} z_t^*(j)z_t^*(j)' - \frac{1}{i^2}\sum_{j=1}^{i} z_t^*(j)\sum_{j=1}^{i} z_t^*(j)'$ when $i > i^*$[9]. i^* is a fixed

[9]To avoid the risk of near-singular sample covariance matrix, one can add $\frac{\epsilon}{i}I_K$ to A in the $i+1^{\text{th}}$ draw where ϵ is a small positive number (e.g. 1e-6).

threshold to avoid unstable sample covariance matrix of z_t^* when i is small. The scalar w_i^2 is updated according to $\log(w_{i+1}) = \log(w_i) + \frac{c}{d_i}(p_i - p^*)$ where p_i is the MH acceptance probability in the i^{th} draw of z_t^* and p^* is the target acceptance probability. The scalar c for updating w is determined as:

$$c = \frac{1}{Kp^*(1-p^*)} + \left(1 - \frac{1}{K}\right) \frac{\sqrt{2\pi}\exp\left(\frac{\alpha_w^2}{2}\right)}{2\alpha_w} \tag{6}$$

where α_w satisfies $\Phi(-\alpha_w) = \frac{p^*}{2}$. As suggested in Garthwaite et al. (2016), the scalar d_i for updating w is set as $\max(\frac{i}{K}, d^*)$, where d^* is a fixed threshold to avoid that w converges before the sample covariance matrix of z_t^* stabilizes. The update of the scalar w is re-started whenever $\log(w)$ changes more than $\log(3)$ from its value at the start or the most recent re-start in order to reduce the impact of a poor starting value of w. In this paper, I set $p^* = 0.25$, $i^* = 100$ and $d^* = 200$. The configurations of adaptive MH steps for μ, a, v and β_0 are similar to the example of z_t^*.

3.2 ASIS Boosting

In my experiments, the sampling quality of parameters in the LMI model by the Gibbs sampler of Section 3.1 could still be unsatisfactory in certain situations. For example, in the simulation study when a regression coefficient contains occasional structural breaks, draws of the parameter v could converge slowly and lead to extremely low acceptance rate of the MH step. To boost the Gibbs sampler, the *ancillarity-sufficiency interweaving strategy* (ASIS) of Yu and Meng (2011) is adopted.

The ASIS provides a principled way to connect sampling from different parametrizations of a multi-level model and thus allows the sampler to explore the parameter space more efficiently. Details of the ASIS can be found in the original paper of Yu and Meng (2011). Implementing the ASIS for the LMI model is described in Appendix E. With a marginal computation cost (a few seconds per 1,000 draws in applications of this paper), the ASIS

steps noticeably improve the sampling efficiency of the LMI model in my experiments.

4 Simulation Study

A simulation study is conducted to compare the effectiveness of the proposed LMI model with the original MI one. The data generating process (DGP) is a linear regression with 5 coefficients of different time variation patterns:

1. Random walk: $\beta_{1,t} = \sum_{j=1}^{t} u_j$ with $u_j \sim N(0, 0.01)$.

2. Change point: $\beta_{2,t} = 2\, I_{\{t_1 < t \leq t_2\}}$.

3. Mixture of constant parameter, random walk and change point:

$$\beta_{3,t} = \left(\sum_{j=1}^{t} u_j \right) I_{\{t_1 < t \leq t_2\}} + I_{\{t > t_2\}}$$

   with $u_j \sim N(0, 0.01)$.

4. Ones: $\beta_{4,t} = 1$.

5. Zeros: $\beta_{5,t} = 0$.

The regressors are from standard normal distributions. The dependent variable is generated by adding a noise from $N(0, \sigma_0^2)$ where σ_0^2 is calibrated such that the ratio of σ_0^2 to the variance of the dependent variable is 0.1.[10] A sample of 300 data points is simulated from the DGP. For the change point and mixture coefficients, the break points are $t_1 = 100$ and $t_2 = 200$.

When estimating the MI model, the priors of overlapping parameters are the same as in the LMI model. The priors of the transition probabilities in the MI model are Beta$(50, 0.5)$ that favor persistent regimes. In estimation, the dependent variable is taken

[10] In experiments, I also simulate data sets where the ratio of σ_0^2 to the variance of the dependent variable is 0.5 and 0.8 respectively. The estimation results are qualitatively similar.

to be homoskedastic, i.e. $\sigma_t^2 = \sigma^2$. The burn-in length is 2,000 after which 10,000 posterior draws are kept for analysis. On a standard desktop computer with a 3.0 GHz Intel Core i5 CPU running in MATLAB R2020b, generating 1,000 posterior draws from the MI model requires 370 seconds while the LMI model takes only 74 seconds.

Figure 2 shows the point-wise posterior median and 90% credible set of the regression coefficient β_t by the LMI model, along with the true coefficient value. The model successfully captures the time variations in β_t with its true value covered by the 90% credible set reasonably well. The β_t estimates, both the posterior median and credible set, by the MI model are visually indistinguishable from those by the LMI model and are not shown to save space.

To quantify the accuracy of β_t estimates, Figure 3 compares the point-wise root mean squared error (RMSE) of β_t by the MI and LMI models:

$$\text{RMSE}_{j,t} = \sqrt{\frac{1}{M}\sum_{i=1}^{M}\left(\hat{\beta}_{j,t}^{(i)} - \beta_{j,t}\right)^2} \tag{7}$$

where $\hat{\beta}_{j,t}^{(i)}$ denotes the i^{th} posterior draw of the coefficient from a given model and $\beta_{j,t}$ the true coefficient value for $j = 1, ..., 5$, $t = 1, ..., 300$, and $i = 1, ..., M$. It can be seen in Figure 3 that there are some marginal difference between the RMSE of the two models but overall the estimation accuracy of β_t from the two models is at the same level. For robustness check, such a comparison of RMSEs of β_t is conducted over 20 data series simulated from the DGP and the results are similar to the ones presented.

To compare the dynamic shrinkage exerted by the various models, it is useful to examine the estimated conditional standard deviation (SD) $\sqrt{d_{j,t}v_j^2}$ of the regression coefficients. Figure 4 shows the point-wise posterior mean and 90% credible set of the conditional SD of the regression coefficients by the MI and LMI models. Across the 5 coefficients, the point-wise posterior mean of the conditional SD from the LMI model tends to be slightly larger than that from the MI model, while the point-wise 95^{th} percentile of the conditional SD

from the LMI model tends to be marginally lower, leading to slightly narrower point-wise 90% credible sets of the conditional SD than the MI model. Nevertheless, such differences between the LMI and MI models are qualitatively minor. The time variation patterns of the estimated conditional SD from these two models are close. The overall posterior uncertainty of the conditional SD estimates as judged by the width of these credible sets is at about the same level between these two models.

In summary, introducing the "continuous" mixture indicators as in the LMI model to perform dynamic shrinkage significantly reduces the computation time relative to the original MI model with discrete mixture indicators even in such a small scale regression. On the other hand, the estimation accuracy of the time varying coefficients as well as the magnitude of dynamic shrinkage, as judged by the estimated conditional SD of the time varying coefficients, is overall comparable between the MI and LMI models, supporting the LMI model as a viable alternative to the original MI model.

Estimation of the LMI model relies heavily on adaptive MH steps. For the simulated data, the acceptance rates of the MH steps for μ, a, v, β_0 and z_t^* over $t = 1, ..., 300$ are all between 0.22 and 0.27 and are close to the target value of 0.25. The scaling parameters of the random walk proposals appear to be stabilized. Figure 5 shows the point-wise inefficiency factor (IF) of estimated β_t from the LMI model. The IF is computed based on the initial monotone sequence method of Geyer (1992) with a smaller IF value implying less correlated and hence better mixed posterior draws. It can be seen that posterior draws of β_t are well mixed with IFs generally below 20 over all $t = 1, ..., 300$.

5 Empirical Illustration

The proposed LMI model is applied to two empirical exercises. Section 5.1 provides a study of predicting the equity premium by the set of variables analyzed in Welch and Goyal

(2008), while Section 5.2 predicts the inflation rate by a number of predictors suggested in the literature. Similar predictive exercises were studied in Kalli and Griffin (2014). In both applications, the number of regressors is relatively large such that estimating these TVP regression models would be computationally prohibitive if the original MI model is applied.

Out-of-sample forecasts are used to test the performance of the LMI model against two alternative TVP models with dynamic shrinkage features. The first alternative model is a restricted version of the MI model (*RMI* hereafter) that places conventional spike-and-slab distributions for the coefficient innovations but limits the number of scenarios the mixture indicators can take in order to be computationally practical. The second alternative model is the dynamic horseshoe (*DHS* hereafter) model of Kowal et al. (2019) that applies a dynamic version of the popular horseshoe prior to the coefficient innovations and has shown good performance in a number of subsequent studies (Hauzenberger et al. (2020), Huber and Pfarrhofer (2021)). Appendix F describes the details of these two alternative models.

5.1 Equity Premium

In the equity premium prediction model, the dependent variable is the value-weighted quarterly return of the S&P500 index minus the corresponding risk free rate. The set of predictors include stock characteristics, interest rates and other macroeconomic indicators. A list of the predictors plus brief descriptions can be found in Appendix G whose detailed descriptions can be found in the original paper of Welch and Goyal (2008). Including the intercept and a first-order AR lag, there are a total of 14 regressors in the model that predicts the equity premium one quarter ahead. The data is kindly provided by Amit Goyal in his website[11]. The data sample runs from Q1 1947 to Q4 2020 with a total of

[11]The web address is https://sites.google.com/view/agoyal145/?redirpath=/.

296 observations. The SV model of Equation (3) is applied to the conditional variance of the dependent variable. In estimation, all non-constant regressors are normalized by subtracting their sample means and dividing by their sample standard deviations.

A total of 10,000 posterior draws from the LMI model are collected for analysis after a burn-in length of 2,000. The acceptance rates of MH steps are between 0.23 to 0.26 and are close to the target value of 0.25. Posterior draws of the model parameters mix reasonably well. For example, the inefficiency factors of the log transformations of the parameters $\log(v^2)$ and $\log(a^2)$ are capped at 56. The running time for 1,000 draws is 116 seconds. In contrast, with 14 regressors estimating the original MI model needs to evaluate $2^{14} = 16,384$ combinatorial scenarios when sampling the 0/1 mixture indicator at each time point t and would require over 66 hours to produce 1,000 draws.

Besides the intercept, there are three predictors that appear significant throughout the data sample: *dividend price ratio* (positive sign), *long term yield* (negative sign) and *investment-to-capital ratio* (negative sign). The two predictors *term spread* and *default return spread* were significantly positive at the beginning of the data sample but gradually become insignificant. Figure 6 provides the point-wise posterior median and 90% credible set of the coefficients of these five predictors. Coefficients of the other predictors including the autoregressive lag are generally flat with the value of zero around the middle of their credit sets.

Recursive predictions are conducted over a 10-year period from Q1 2011 to Q4 2020. In each forecast, an one-quarter-ahead prediction of the equity premium is generated to compute a predictive likelihood that integrates out all model parameters and latent variables. For both the LMI and alternative TVP models, β_t is analytically integrated out by a Kalman filter step at each MCMC draw when generating predictions in order to improve their numerical stability. Since the non-constant regressors are normalized in estimation, the input variables to each prediction are recursively normalized by the sample moments

of the estimation sample only.

Predictions of the LMI and alternative models are compared by the cumulative difference in their log one-quarter-ahead predictive likelihoods that shows the entire evolution of each model's relative performance over the prediction sample. See Geweke and Amisano (2010) for a review of Bayesian predictive analysis. The upper panel of Figure 8 plots the sequence of the cumulative log predictive likelihood of the LMI model minus that of the RMI and DHS models. It can be seen that, except for occasional setbacks, the LMI model steadily accumulates gains in predictive likelihoods over the RMI and DHS models. At the end of the prediction sample, the cumulative log predictive likelihood of the LMI model is 18.6 over the RMI model and is 7.4 over the DHS model, which would be interpreted as "decisive" evidence for the LMI model based on the interpretation of Bayes factor scale in Kass and Raftery (1995).[12] As a commonly used gauge in prediction studies, a Diebold-Mariano test (Diebold and Mariano (1995)) is conducted for the average difference of log predictive likelihoods between the LMI and the RMI and DHS models. The resulting test statistics are 5.7 and 5.0 for the LMI-RMI and LMI-DHS pairs respectively and strongly favor the LMI model.

5.2 Inflation Rate

The second application is predicting the quarter-to-quarter change of the U.S. inflation rate as measured by the quarterly change of log GDP deflator. A total of 20 exogenous predictors are considered including real activity variables, interest rates and other macroeconomic indicators. A list of the exogenous variables and their descriptions are provided in Appendix

[12] An exercise is also conducted to estimate a small predictive regression model with only the intercept, AR lag and dividend price ratio by the LMI and unrestricted MI models. The out-of-sample predictive likelihoods from the two models are close with the maximum cumulative difference of 1.2 over the prediction sample.

H. Along with the intercept and an AR(1) lag, the one-quarter-ahead prediction model includes 22 regressors.

Data on the S&P500 index is from Robert Shiller's website[13]. Data on all other variables are from the FRED database of the U.S. Federal Reserve Bank of St. Louis. The data sample is quarterly from Q2 1966 to Q1 2021 with a total of 220 observations. The quarterly value of monthly variables are constructed as the monthly averages within each quarter. Similar to the equity premium prediction model, the SV model of Equation (3) is estimated for the conditional variance of the dependent variable. Non-constant regressors are normalized in estimation.

The LMI model is estimated with 10,000 posterior draws after a burn-in of 2,000. With a target of 0.25 for MH acceptance rates, the actual MH acceptance rates are between 0.24 to 0.26. Posterior draws of the model parameters show reasonable mixing behavior. For example, the inefficiency factors of the log transformations of the parameters $\log(v^2)$ and $\log(a^2)$ are less than 26. In terms of computation time, producing 1,000 draws from the LMI model takes 139 seconds. In contrast, directly applying the original MI model would require over 16 hours to produce a single draw given the $2^{22} = 4,194,304$ combinatorial scenarios of the 0/1 mixture indicator at each time point t.

Among the regressors, only the AR lag is consistently significant. The predictors *GDP*, *employee* and *unemployment* are insignificant at the beginning of the sample and gradually become significant. There are also variables such as *investment*, *expenditure* and *producer price* that go from significant to insignificant. Figure 7 plots the point-wise posterior median and 90% credible set of the coefficients of the AR lag, *GDP*, *employee* and *unemployment*.

Similar to the equity premium prediction model, the LMI model is compared to alternative models in out-of-sample predictions of inflation rate. Recursive predictions are conducted over a 10-year period from Q1 2011 to Q1 2021. The methodology of generating

[13]The web address is http://www.econ.yale.edu/ shiller/data.htm.

predictive likelihoods for inflation rate is identical to that in the equity premium prediction exercise. The lower panel of Figure 8 shows the cumulative log predictive likelihood of the LMI model minus that of the RMI and DHS models. For this application, the negative impact of imposing restrictions on the original MI model happens to be less severe. The difference of the cumulative log predictive likelihood of the LMI model over the RMI model is relatively small and is 2.1 at the end of the prediction sample. On the other hand, the advantage of the LMI model over the DHS model is significant. At the end of the prediction sample, the cumulative log predictive likelihood of the LMI model is 6.9 over the DHS model and is "decisive" evidence favoring the LMI model based on the interpretation of Bayes factor scale in Kass and Raftery (1995). A Diebold-Mariano test of the average difference of log predictive likelihoods between the LMI and DHS models returns a statistic of 3.0 and is highly significant.

6 Conclusion

This paper proposes a new MI model that uses a logistic function of a latent continuous variable to mimic the behavior of the 0/1 mixture indicator in the original MI model of Giordani and Kohn (2008) in order to allow flexible time variations in regression coefficients while imposing dynamic shrinkage. The resulting model retains the attractive shrinkage feature of the original MI model while enjoying a large computational advantage, thus opening the door of applying spike-and-slab-like priors to perform dynamic shrinkage for larger TVP regression problems with a manageable computation cost.

An MCMC algorithm is developed to estimate the LMI model that applies the adaptive MH method of Garthwaite et al. (2016) and boosts the Gibbs sampler by the ASIS of Yu and Meng (2011). In a simulation study with a relatively small number of regressors, the LMI model produces estimates of the regression coefficients that are at the same level

of accuracy as the original MI model. The LMI model also works well in two empirical exercises with a relatively large number of regressors where applying the original MI model is impractical and is shown to produce larger predictive likelihoods than popular alternative dynamic shrinkage models such as a restricted version of the original MI model and the DHS model of Kowal et al. (2019).

The effectiveness of the adaptive MH steps when the number of regressors is extremely large (e.g. in the order of hundreds or even more) might be a potential concern. It is conceivable that a practical strategy of estimating the LMI model in this case could be to divide the regressors into blocks of more manageable size, preferably based on the correlation between regressors, and apply the adaptive MH steps to iterate sampling model parameters over the blocks. Further study of the model sampler's large-dimension scalability could be an interesting future research direction.

References

Belmonte, M., G. Koop, and D. Korobolis (2014). Hierarchical shrinkage in time-varying parameter models. *Journal of Forecasting 33*, 80–94.

Bitto, A. and S. Fruhwirth-Schnatter (2019). Achieving shrinkage in a time-varying parameter model framework. *Journal of Econometrics 210*, 75–97.

Cadonna, A., S. Fruhwirth-Schnatter, and P. Knaus (2020). Triple the gamma - a unifying shrinkage prior for variance and variable selection in sparse state space and tvp models. *Econometrics 8*(2), 20.

Carriero, A., T. Clark, and M. Marcellino (2019). Large bayesian vector autoregressions with stochastic volatility and non-conjugate priors. *Journal of Econometrics 212*(1), 137–154.

Carvalho, C., N. Polson, and J. Scott (2009). Handling sparsity via the horseshoe. In D. van Dyk and M. Welling (Eds.), *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics*, Volume 5. JMLR: W&CP.

Carvalho, C., N. Polson, and J. Scott (2010). The horseshoe estimator for sparse signals. *Biometrika 97*, 465–480.

Chan, J., E. Eisenstat, and R. Strachan (2020). Reducing the state space dimension in a large tvp-var. *Journal of Econometrics 218*, 105–118.

Chan, J., G. Koop, R. Leon-Gonzalez, and R. Strachan (2012). Time varying dimension models. *Journal of Business and Economic Statistics 30*(3), 358–367.

Chang, Y., Y. Choi, and J. Park (2017). A new approach to model regime switching. *Journal of Econometrics 196*(1), 127–143.

Chib, S. (1998). Estimation and comparison of multiple change-point models. *Journal of Econometrics 86*(2), 221–241.

Diebold, F. and R. Mariano (1995). Comparing predictive accuracy. *Journal of Business and Economics Statistics 13*, 253–263.

Dufays, A., Z. Li, J. Rombouts, and Y. Song (2021). Sparse change-point var models. *Journal of Applied Econometrics*, forthcoming.

Durbin, J. and S. Koopman (2002). A simple and efficient simulation smoother for state space time series analysis. *Biometrika 89*, 603–615.

Fruhwirth-Schnatter, S. and H. Wagner (2010). Stochastic model specification search for gaussian and partially non-gaussian state space models. *Journal of Econometrics 154*, 85–100.

Garthwaite, P., Y. Fan, and S. Sisson (2016). Adaptive optimal scaling of metropolis-hastings algorithms using the robbins-monro process. *Communications in Statistics - Theory and Methods 45*(17), 5098–5111.

George, E. and R. McCulloch (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association 88*, 881–889.

Gerlach, R., C. Carter, and R. Kohn (2000). Efficient bayesian inference for dynamic mixture models. *Journal of the American Statistical Association 95*, 819–828.

Geweke, J. and G. Amisano (2010). Comparing and evaluating bayesian predictive distributions of asset returns. *International Journal of Forecasting 26*, 216–230.

Geyer, C. (1992). Practical markov chain monte carlo. *Statistical Science 7*, 473–483.

Giordani, P. and R. Kohn (2008). Efficient bayesian inference for multiple change-point and mixture innovation models. *Journal of Business and Economic Statistics 26*, 66–77.

Hamilton, J. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica 57*(2), 357–384.

Hauzenberger, N., F. Huber, and G. Koop (2020). Dynamic shrinkage priors for large time-varying parameter regressions using scalable markov chain monte carlo methods. arXiv:2005.03906v1 [econ.EM].

Hoeting, J., D. Madigan, A. Raftery, and C. Volinsky (1999). Bayesian model averaging: a tutorial. *Statistical Science 14*, 382–417.

Huber, F., G. Kastner, and M. Feldkircher (2019). Should i stay or should i go? a latent threshold approach to large-scale mixture innovation models. *Journal of Applied Econometrics 34*(5), 621–640.

Huber, F. and M. Pfarrhofer (2021). Dynamic shrinkage in time-varying parameter stochastic volatility in mean models. *Journal of Applied Econometrics 36*, 262–270.

Ishwaran, H. and J. Rao (2005). Spike and slab variable selection: frequentist and bayesian strategies. *Annals of Statistics 33*, 730–773.

Kalli, M. and J. Griffin (2014). Time-varying sparsity in dynamic regression models. *Journal of Econometrics 178*, 779–793.

Kass, R. and A. Raftery (1995). Bayes factors. *Journal of the American Statistical Association 90*, 773–795.

Kastner, G. and S. Fruhwirth-Schnatter (2014). Ancillarity-sufficiency interweaving strategy (asis) for boosting mcmc estimation of stochastic volatility models. *Computational Statistics and Data Analysis 76*, 408–423.

Koop, G. and S. Potter (2007). Estimation and forecasting in models with multiple breaks. *Review of Economic Studies 74*(3), 763–789.

Kowal, D., D. Matteson, and D. Ruppert (2019). Dynamic shrinkage processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 81*, 781–804.

Kuo, L. and B. Mallick (1998). Variable selection for regression models. *Sankhya: Series B 60*, 65–81.

Lopes, H., R. McCulloch, and R. Tsay (2021). Parsimony inducing priors for large scale state space models. *Journal of Econometrics*, forthcoming.

Maheu, J. and S. Gordon (2008). Learning, forecasting and structural breaks. *Journal of Applied Econometrics 23*, 553–583.

Makalic, E. and D. Schmidt (2016). A simple sampler for the horseshoe estimator. *IEEE Signal Processing Letters 23*(1), 179–182.

McCausland, W., S. Miller, and D. Pelletier (2011). Simulation smoothing for state-space models: A computational efficiency analysis. *Computational Statistics and Data Analysis 55*(1), 199–212.

Mitchell, T. and J. Beauchamp (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association 83*, 1023–1036.

Nakajima, J. and M. West (2013). Bayesian analysis of latent threshold dynamic models. *Journal of Business and Economic Statistics 31*, 151–164.

Primiceri, G. (2005). Time varying structural autoregressions and monetary policy. *Review of Economic Studies 72*(3), 821–852.

Rue, H. (2001). Fast sampling of gaussian markov random fields. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63*, 325–338.

Terasvirta, T. (1998). Modelling economic relationships with smooth transition regressions. In A. Ullah and D. Giles (Eds.), *Handbook of Applied Economic Statistics*, pp. 507–552. New York: Marcel Dekker.

Tong, H. (1990). *Non-Linear Time Series: A Dynamical Systems Approach*. Oxford: Oxford University Press.

Uribe, P. and H. Lopes (2020). Dynamic sparsity on dynamic regression models. arXiv:2009.14131v1 [stat.ME].

Welch, I. and A. Goyal (2008). A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies 21*(4), 1455–1508.

Yu, Y. and X. Meng (2011). To center or not to center: That is not the question - an ancillarity-sufficiency interweaving strategy (asis) for boosting mcmc efficiency. *Journal of Computational and Graphical Statistics 20*(3), 531–570.

Appendix

A Hyper-Parameters of Horseshoe Prior

B Estimating SV Model

C Sampling the Latent Variable z_t^*

D Computing the Model Likelihood Function

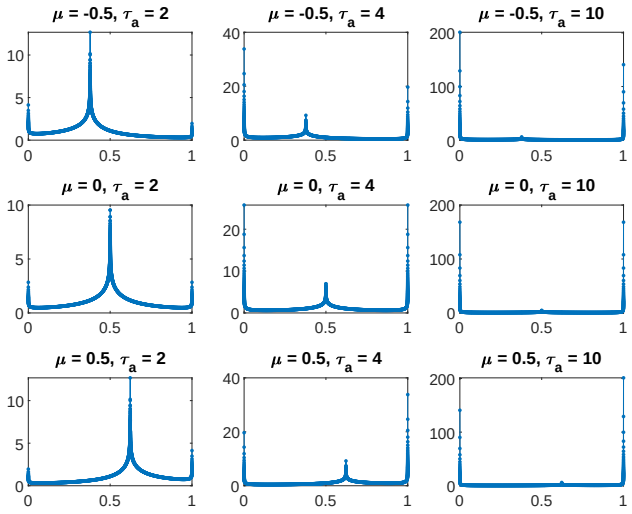
E ASIS for the LMI Model

F Alternative TVP Models with Dynamic Shrinkage

G Equity Premium Data

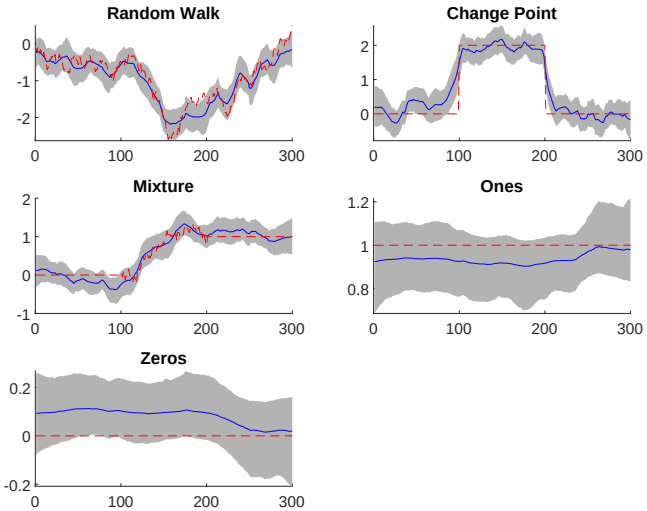
H Inflation Rate Data

Figure 1: Conditional Density of the Mixture Indicator d_t



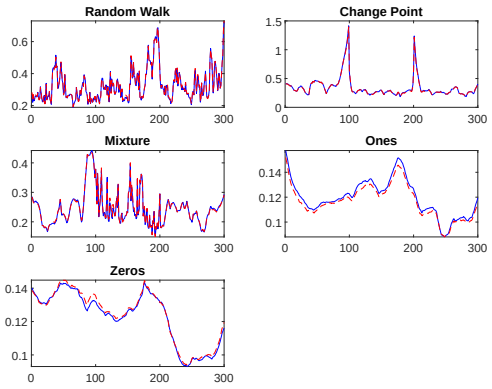
Note: The figure shows the conditional density $p(d_t|\mu, \tau_a)$ for the mixture indicator $d_t \in (0, 1)$ over different values of the conditional mean parameter μ and the hyper-parameter τ_a of the conditional variance parameter a^2 where $a^2 \sim G(0.5, 2\tau_a)$.

Figure 2: Estimate of β_t: Simulation Study



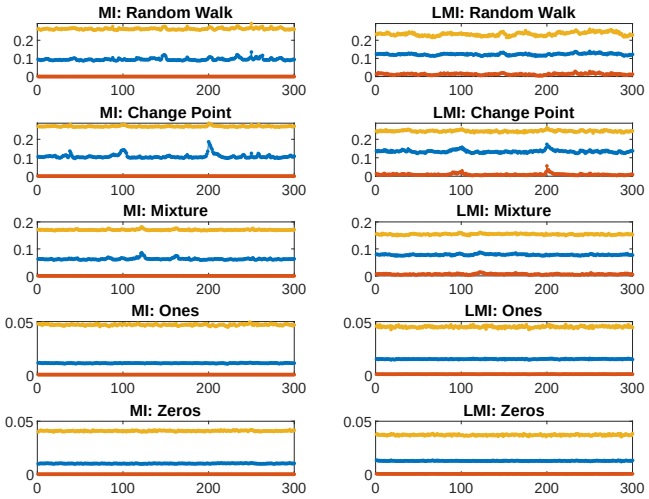
Note: The figure shows the point-wise posterior median (solid line) and 90% credible set (gray area) of estimated coefficients β_t for $t = 1, ..., 300$ by the LMI model. The dashed line is the true coefficient.

Figure 3: Comparing Root Mean Squared Error of β_t Estimates: Simulation Study



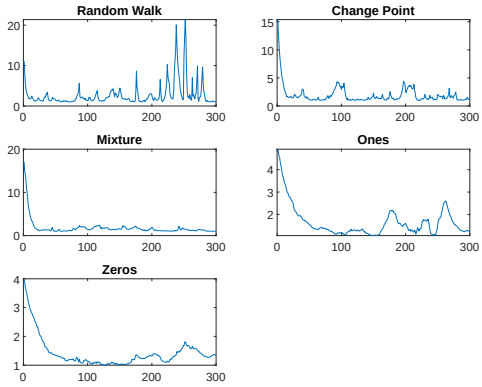
Note: The figure compares the point-wise root mean squared error (Equation (7)) of estimated coefficients β_t for $t = 1, ..., 300$ by the LMI and MI models. The solid line is the root mean squared error by the LMI model while the dashed line is by the MI model.

Figure 4: Estimated Conditional Standard Deviations of β_t: Simulation Study



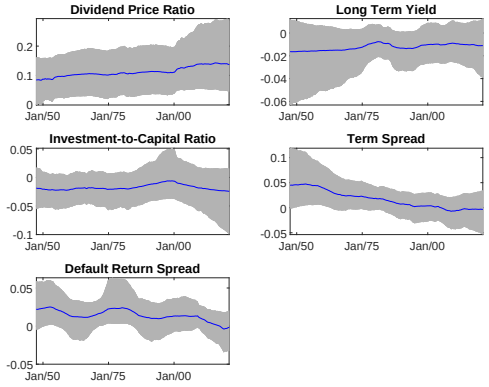
Note: The figure compares the estimated conditional standard deviations of the coefficients β_t for $t = 1, ..., 300$ by the MI and LMI models. Each panel plots the point-wise posterior mean (the central line) and 90% credible set of the conditional standard deviations by one of the models under study.

Figure 5: Point-Wise Inefficiency Factor of β_t Estimates: Simulation Study



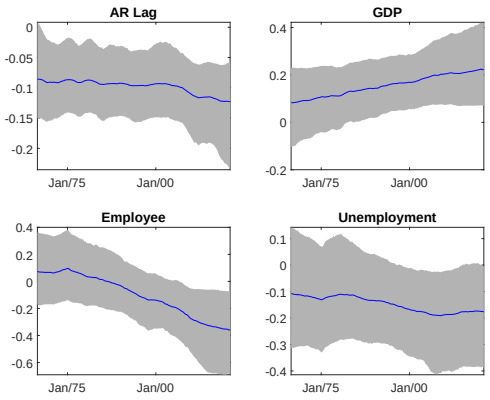
Note: The figure shows the point-wise inefficiency factor of estimated coefficients β_t by the LMI model. A smaller inefficiency factor indicates less correlated posterior draws and better mixing behavior.

Figure 6: Estimate of Selected β_t: Equity Premium Prediction



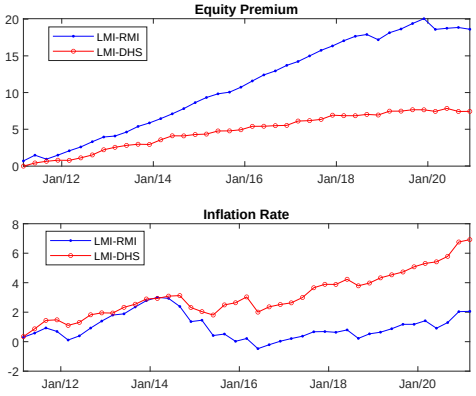
Note: The figure shows the point-wise posterior median (solid line) and 90% credible set (gray area) of estimated coefficient β_t by the LMI model for the 5 regressors in the equity premium prediction model that are either significant throughout the data sample (dividend price ratio, long term yield, investment-to-capital ratio) or are significant at the beginning of the data sample (term spread, default return spread). Description of the regressors can be found in Appendix G.

Figure 7: Estimate of Selected β_t: Inflation Rate Prediction



Note: The figure shows the point-wise posterior median (solid line) and 90% credible set (gray area) of estimated coefficient β_t by the LMI model for the 4 regressors in the inflation rate prediction model that are either significant throughout the data sample (AR lag) or are significant at the end of the data sample (GDP, employee, unemployment). Description of the regressors can be found in Appendix H.

Figure 8: Comparing Cumulative Difference of Log Predictive Likelihoods



Note: Abbreviate the cumulative difference of log predictive likelihoods as *CDL*. The figure shows the CDL of the LMI model versus the two alternative models: RMI (dot marker) and DHS (circle marker) in predicting the equity premium and the inflation rate respectively. In all panels a positive CDL value favors the LMI model.