Econometrics and Statistics

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

--Manuscript Draft--

Manuscript Number:	ECOSTA-D-22-00151R1
Article Type:	Part A: Econometrics
Keywords:	TVP; Dynamic Shrinkage; Spike-and-Slab Prior; Mixture Innovation
Corresponding Author:	Zhongfang He CANADA
First Author:	Zhongfang He
Order of Authors:	Zhongfang He
Abstract:	The mixture innovation (MI) model places spike-and-slab mixture distributions on the innovations of time-varying regression coefficients and permits flexible time variation patterns while allowing for dynamic shrinkage. Despite its appeal, the standard algorithm requires O(2 K n) operations to draw the mixture indicators in the MI model for a data set with n observations and K regressors and is computationally prohibitive when K is even modestly large. As an alternative, a new specification of the MI model is proposed 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 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. A simulation study shows 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.
Response to Reviewers:	Reply to the editor's comments can be found in the file "Reply_Editor" submitted as the cover letter. Point-to-point replies to the two referees can be found in the files "Reply_R1" (ECOSTA-D-22-00151-ref-au) and "Reply_R2" (report-ECOSTA).

Thanks for the opportunity to revise and resubmit my paper. The draft has been updated by incorporating the referees' comments. A reply to each of the two referee reports is attached in the re-submission which contains point-to-point replies (R1=ECOSTA-D-22-00151-ref-au, R2=report-ECOSTA).

Following your comments, the reference in the abstract is removed. All tables and figures are moved from the end to the main text. Best efforts have been made to proofread the draft for grammatical errors. No author names and affiliations are shown in the updated draft as well as the replies.

Thanks for your invaluable comments.

The draft has been updated following the referees' comments. Among the changes, the major ones are the following:

1. The theoretical discussion of the distribution of the mixture indicator d_t in the proposed model has been updated (Section 2.2 starting from page 8). In particular, a discussion of the distribution of d_t conditional on d_{t-1} and the autoregressive coefficient ρ is added.

2. An appendix (Appendix A on page 30) is added to include additional results of the simulation study that compares the width of the pointwise credible set of the coefficients β_t by the original MI model and the proposed one.

Below I reply to each of the points you raised.

1. *The manuscript repeatedly advertises $\mathcal{O}(n)$ scalability instead of $\mathcal{O}(2^K n)$ for certain spike-and-slab alternatives. First, I do not see how this is possible: the computing time must depend on the number of covariates K, even if it is linear. Second, it is widely common in spike-and-slab models to update one indicator at a time from its full conditional distribution (i.e., given the remaining indicators), which is linear in K. Such a stochastic search must still explore the 2^K space, but would seem to achieve the same computational scalability as the proposed approach. Why is this not considered or discussed as an alternative here?*

   Reply: Here the focus is on comparing the cost of drawing the mixture indicators in the MI model and the proposed logistic alternative. Computing the model likelihood does involve operations of the data matrix of regressors and hence its ***exact*** computing cost depends on K. However such computation costs are the same in both models and would be irrelevant when the aim is comparing the order of magnitude of computation complexity. The $\mathcal{O}(n)$ or $\mathcal{O}(2^K n)$ cost concerns the number of times the model likelihood needs to be evaluated.

As in your comment, it is possible to cycle through individual elements of the indicators to avoid the combinatorial complexity in block update of the indicators (i.e. update $d_{j,t}$ over $j = 1, ..., K$ instead of jointly update $d_t = [d_{1,t} \ ... \ d_{K,t}]'$). However such an approach is usually not used for estimating the MI model in the literature considering that there are a large number of nK elements in the indicators. It is a matter of sampling $d = \{d_t\}_{t=1}^n$ in nK blocks (one-element-at-a-time sampler) or in n blocks that jointly update each d_t. Updating one element at a time leads to loss of sampling efficiency when K grows. The usual practice would be to keep block updates while restricting the possible scenarios of the indicators or reducing the dimension of the indicators. For instance, Chan et al. (2012) is an example that develops certain restricted versions of the MI model to circumvent the computation difficulty of block updating the indicators. In terms of computing cost, the one-element-at-a-time sampler would need to evaluate the model likelihood $2nK$ times which is K times of the computing cost of the proposed approach.

In view of your comment, a footnote is added (footnote 1, page 3) to discuss the one-element-at-a-time sampling approach.

2. *The simulation study compares the proposed approach (LMI) to a spike-and-slab analog (MI). However, the forecasting comparison in the next section instead uses a restricted version of the MI (RMI). I dont see why this restricted version is necessary, especially in light of the comments above.*

Reply: The data generating process in the simulation study is designed to have a small number of regressors ($K = 5$) such that estimating the original MI model with block update of the indicators is practical (though still the running time is inconveniently long). In the empirical examples, K is too large for estimating the original MI model and hence the RMI is used instead. Estimating the unrestricted MI model with

an one-element-at-a-time sampler loses sampling efficiency as well as requiring K times more computing cost than the proposed approach as explained in the reply to comment 1.

3. *Previous work on dynamic shrinkage has shown that gains from these efforts are often in uncertainty quantification (i.e., narrower credible intervals that maintain nominal coverage), rather than point estimation or forecasting. Given that the proposed method does advertise dynamic shrinkage, it would improve the simulation study to report mean credible interval widths and empirical coverage for the regression coefficients, and to compare them to dynamic shrinkage alternatives.*

   Reply: In the updated draft, an appendix (Appendix A on page 30) is added that compares the mean width of point-wise 90% and 95% credible sets for the regression coefficients in the simulation study under the MI and LMI models. As shown in Table A1 (page 30) of the appendix, the mean width averaged over time points $t = 1, ..., n$ is very close between the two models.

   As for empirical coverage, it seems still an active research area to establish the relation between Bayesian credible set and nominal frequentist coverage probability of latent variable models (e.g. Martin and Ning (2020) and references therein). For example, for the change point coefficient of the simulation study, TVP models tend to be unable to exactly identify the change points. The empirical coverage at such change points would be substantially below the nominal probability. As such I am not sure I have fully grasped the idea of empirical coverage for the TVP setting in the present paper. Given these considerations, discussion of empirical coverage is not included in the main text to avoid digression from the main point of the paper. However it is shown in Table 1 of the reply that the empirical coverage of β_t by the MI and LMI models, averaged over time points $t = 1, ..., n$, is very close based on 20 simulated data samples that keep the true β_t fixed, confirming

the findings in the paper that the outputs of the proposed model are comparable to the original MI model.

Table 1: Average Empirical Coverage of β_t

	90% Credible Set		95% Credible Set	
	MI	LMI	MI	LMI
Random walk	0.802	0.805	0.908	0.909
Change point	0.863	0.862	0.929	0.929
Mixture	0.902	0.903	0.961	0.959
Ones	0.914	0.909	0.970	0.968
Zeros	0.829	0.832	0.939	0.940

Note: The table compares the average empirical coverage of point-wise 90% and 95% credible sets of the coefficients β_t estimated by the MI and LMI models. The empirical coverage is based on 20 simulated data sets while the average empirical coverage is the average of point-wise empirical coverages over the time points $t = 1, ..., 300$.

4. *The forecasting results seem to favor the proposed approach over the competing methods. However, it is not clear to me why the results should be so significant: these methods are estimating very similar models, and the main difference is one of computing. What is the explanation for such dramatic improvements in performance? Can this be explained better through the simulation study?*

   Reply: The empirical examples compare the forecast performance of the proposed LMI model with a restricted variant of the MI model and the dynamic horseshoe model. The difference between these models goes beyond computing and is mainly on how the process variances of the time-varying coefficients are modeled. Such a difference leads to the different shrinkage abilities of these models for regularizing the time-varying coefficients which in turn affects their forecasting results. The improved forecast performance over the dynamic horseshoe model shows the benefit of the spike-and-slab-like distribution of the process variances in the proposed model, while the improvement over the re-

stricted MI model shows the benefit of allowing greater flexibility for the mixture indicators.

In the equity premium exercise (footnote 14 of page 24), a smaller model with only 3 regressors was estimated by both the proposed model and the unrestricted MI model. Here the main difference between the two models is computing. Such a computing difference does not cause much difference in their forecast performance.

5. *A related concern is the sensitivity to hyperparameters, which is a notorious challenge for spike-and-slab priors. Does that issue arise for this prior?*

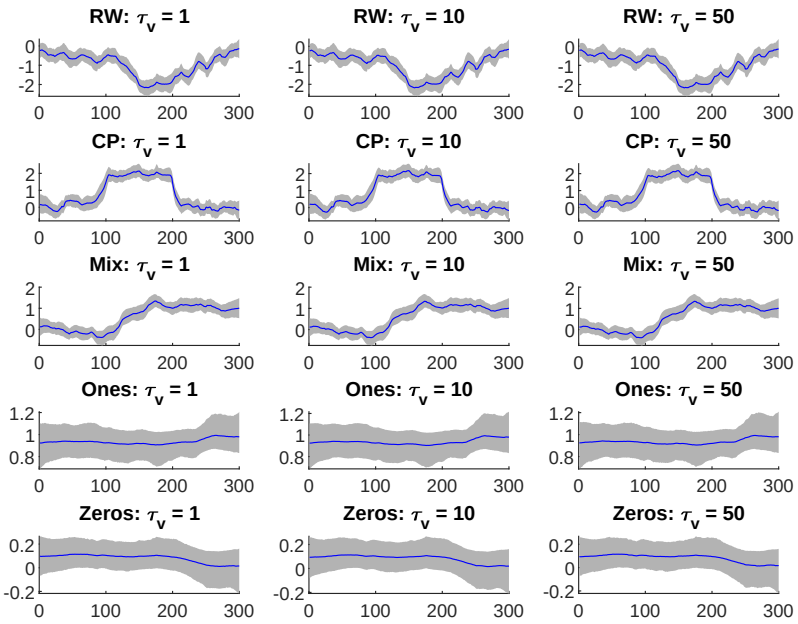
   Reply: A major hyper-parameter is the one for the conditional variance a^2 of the latent variable z_t. The prior for a^2 is a gamma distribution $G(0.5, 2\tau_a)$. The hyper-parameter τ_a is crucial for achieving desirable shrinkage as described in the theoretical discussion of Section 2.2 which guides the choice of τ_a in the simulation and empirical exercises.

   Also a relatively tight Gaussian prior is placed for the long-run mean μ of the latent variable z_t. A diffuse prior for μ tends to result in slower mixing of posterior draws.

   Other hyper-parameters are less influential when their values are in reasonable ranges. The prior for the invariant part of the process variance is a gamma distribution $v^2 \sim G(0.5, 2\tau_v)$. The present paper sets the hyper-parameter $\tau_v = 10$. Changing the value of τ_v to be 1 or 50 has little impact on the results. Figure 1 in the reply shows the estimated coefficients β_t under $\tau_v = 1$, 10, 50 for the simulation study. It can be seen that there is little difference in the estimated β_t when τ_v is in a reasonable range.

6. *Given the lack of dynamics in Figure 6, it would be useful to include a non-dynamic Bayesian regression model for benchmarking.*

Figure 1: Estimate of β_t Under Different Values of τ_v



Note: The figure shows the point-wise median (solid line) and 90% credible set (gray area) of coefficients β_t in the simulation study under different values of the hyper-parameter τ_v.

Reply: Following your suggestion, the estimated coefficients of a conventional regression are included as the benchmark (Figure 7 of page 24, Figure 9 of page 28).

7. *Figure 1 shows an unusual trend in that the mixture indicator is trimodal: zero, one, and some other number. This behavior seems to differ from usual spike-and-slab structures. How should one interpret the peak between zero and one, and is it useful for modeling?*

   Reply: The peak in the middle is due to the Gaussian specification of the latent variable z_t and the logistic transformation. Intuitively, being a Gaussian variable, z_t has relatively large probability mass around its mean μ when its variance is small and hence the density of the mixture indicator $d_t = \frac{1}{1+\exp(-z_t)}$ could show a spike around $\frac{1}{1+\exp(-\mu)}$ which lies in the middle of 0 and 1 when μ is of a modest value. As the variance of z_t increases, the probability mass of z_t becomes less concentrated around its mean and the middle peak in the density of d_t is diminished as seen in Figure 1 (page 10).

8. *Inefficiency factors (Figure 5) are largest when t is small. Is there a way to improve the MCMC efficiency for these time points?*

   Reply: The pattern that inefficiency factors are larger for small t may likely be due to the random walk specification of the coefficients which makes exploring the initial value (i.e. $t = 0$) and hence those at small t less efficient. Integrating out the initial value of the coefficients might be possible to reduce this relative inefficiency. However, considering that the inefficiency factors at small t are already reasonably small in the present paper, I would leave the exploration of improving the sampling efficiency at small t to future works.

9. *It is noted that three predictors appear significant (page 18). Can this be made more precise?*

   Reply: To be more precise, the text has be rewritten as *"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) as the value of zero is consistently outside or near the boundaries of the point-wise 90% credible set of their coefficients." (2nd paragraph of page 23).

References

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.

Martin, R. and B. Ning (2020). Empirical priors and coverage of posterior credible sets in a sparse normal mean model. *Sankhya A 82*, 477–498.

Thanks for your invaluable comments.

The draft has been updated following the referees' comments. Among the changes, the major ones are the following:

1. The theoretical discussion of the distribution of the mixture indicator d_t in the proposed model has been updated (Section 2.2 starting from page 8). In particular, a discussion of the distribution of d_t conditional on d_{t-1} and the autoregressive coefficient ρ is added.

2. An appendix (Appendix A on page 30) is added to include additional results of the simulation study that compares the width of the pointwise credible set of the coefficients β_t by the original MI model and the proposed one.

Below I reply to each of the points you raised.

1. *Time-varying parameter regressions are widely used, and increasingly in applications with a large number of predictors. Therefore, I think theres value in developing dynamic shrinkage priors that are not too computationally intensive to implement. The new prior is motivated as a computationally feasible alternative to the mixture innovation model by using a continuous mixture rather than a discrete mixture. As such, this new prior is closer to the class of continuous scale mixtures shrinkage priors than the mixture innovation model. (For example, the new prior does not retain the nice interpretation of the mixture innovation model as Bayesian model averaging over a large number of specifications). It would therefore be better to reposition the new prior as a member of the class of continuous mixtures shrinkage prior.*

2. *The idea of using a continuous mixture rather than a discrete mixture to speed up computation has been around for more than a decade (e.g., Griffin and Brown, 2010; Polson and Scott, 2010). And there are by now a large literature on dynamic shrinkage in TVP settings using continuous scale mixture priors. Other continuous scale mixture priors are*

certainly computationally feasible for settings considered in this paper. Hence, the paper would need to do a better job to motivate the new prior. For instance, Im not able to find any theoretical motivation/discussion on why the new prior is better than other continuous dynamic shrinkage priors. Such information would help the readers better appreciate the contribution of the paper.

Reply: The basic idea of the proposed approach is to approximate the discrete mixture indicator of a spike-and-slab distribution in the MI model by a logistic function of a latent continuous variable for computational efficiency. Depending on how well the logistic function approximates the 0/1 indicator (discussed in Section 2.2 of the paper), the difference between the other continuous dynamic shrinkage priors and the proposed model would essentially boil down to their difference with the spike-and-slab distribution. Theoretical comparison of the spike-and-slab distribution versus the existing continuous shrinkage priors has been extensively discussed in the literature and is not the main focus of the present paper.

In view of your comments, a more detailed comparison of the LMI and dynamic horseshoe models is conducted in Section 2.3 (last paragraph of page 11) to clarify the similarity and difference of the proposed model to a representative of the other continuous dynamic shrinkage priors.

3. *In addition to the dynamic horseshoe prior cited in the paper, the new prior is also similar to the normal-gamma prior for TVP-VARs in Pruser (2021). In both cases the error variance in the state equation, v_j, follows a gamma distribution. It would therefore be useful to discuss the theoretical differences between the two priors and compare them empirically.*

   Reply: Thanks for pointing me to the paper of Pruser (2021).

   Similar to the present paper, the model in Pruser (2021) specifies the time-varying regression coefficients as random walk process with time-

varying process variance. Using the notation of the present paper, each time-varying coefficient in the model of Pruser (2021) follows $\beta_{j,t} = \beta_{j,t-1} + \eta_{j,t}$ where $\eta_{j,t} \sim N(0, d_{j,t}v_j^2)$, $\sqrt{d_{j,t}} \sim C^+(0,1)$ and $\sqrt{v_j^2} \sim C^+(0,1)$. To the extent that the square of a half Cauchy distribution $C^+(0,1)$ can be written as a scale-mixing gamma distribution, the components of the process variance in Pruser (2021) could be treated as following gamma distributions. However such a model is effectively a static variant of the dynamic horseshoe model of Kowal et al. (2019) which has been compared in detail with the proposed model in the present paper (last paragraph of page 11 as well as the forecast exercises of empirical studies). The dynamic horseshoe model specifies the local component $d_{j,t}$ in the process variance as $d_{j,t} = d_{j,t-1}^{\rho_j}\xi_{j,t}$ with $\sqrt{\xi_{j,t}} \sim C^+(0,1)$ to introduce serial correlation in the local component. It is clear that by setting the autoregressive coefficient $\rho_j = 0$, the dynamic horseshoe model nests the "static" horseshoe model in Pruser (2021).

There is also a "normal-gamma" prior studied in Pruser (2021) as an alternative to its main model of the "static" horseshoe prior. In particular, the global component in this normal-gamma prior is an inverse gamma distribution. Studies such as Polson and Scott (2012) have pointed out the theoretical shortcomings of using the inverse gamma distribution in settings where shrinkage is desirable. Investigations in Pruser (2021) also confirms inadequate shrinkage of the normal gamma prior relative to the horseshoe one.

Given these considerations, the approach in Pruser (2021) is treated as a special case of the dynamic horseshoe model and is not singled out for comparison. A reference to Pruser (2021) is added when discussing the existing literature of dynamic shrinkage priors (line 3 of the last paragraph on page 11).

4. *Theres some discussion on a certain conditional distribution of d_{it} in*

Section 2.2 and I appreciate that. But the case is too special ($\rho = 0$ and so theres no dynamics). It would be useful to work out, for example, the conditional distribution of d_t given d_{t-1}, so that it can be compared to other dynamic shrinkage priors (e.g., dynamic horseshoe). Even though this conditional distribution probably cant be derived analytically, it would still be possible to plot the density for some selected values of d_{t-1} (and other parameters) using numerical integration.

Reply: Thanks for your suggestion. The updated draft adds a discussion of the distribution of d_t conditional on various values of the lag d_{t-1} and the AR coefficient ρ where the conditional variance $a^2 \sim G(0.5, 2\tau_a)$ are integrated out analytically and the long-run mean μ are integrated out numerically (2nd paragraph of page 9). Other parts of the section are also rephrased accordingly. The results are largely consistent with intuition and show that the value of d_t is correlated with its lag d_{t-1} and the AR coefficient ρ. In particular, even when the AR effect is strong ($\rho = 0.9$), there is still an appreciable probability mass of d_t concentrating on the opposite end of d_{t-1} and hence the behavior of parameter change at time t could be different from that at time $t - 1$.

References

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

Polson, N. and J. Scott (2012). On the half-cauchy prior for a global scale parameter. *Bayesian Analysis 7*(4), 887–902.

Pruser, J. (2021). The horseshoe prior for time-varying parameter VARs and monetary policy. *Journal of Economic Dynamics and Control 129*, 104188.

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

Abstract

The mixture innovation (MI) model places a spike-and-slab mixture distribution for the innovations of time-varying regression coefficients and permits flexible time variation patterns while allowing for dynamic shrinkage. Despite its appeal, the standard algorithm requires $\mathcal{O}(2^K n)$ operations to draw the mixture indicators in the MI model for a data set with n observations and K regressors and is computationally prohibitive when K is even modestly large. As an alternative, a new specification of the MI model is proposed 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. A simulation study shows 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

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 conventional 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)). To overcome this sampling problem, Gerlach et al. (2000) (*GCK* hereafter) develops a sampler that draws the mixture indicator d by marginalizing over η in efficient $\mathcal{O}(n)$ operations and has become the standard algorithm in the literature (e.g. Giordani et al. (2007), Uribe and Lopes (2020)). 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)).[1]

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

[1]Another alternative is the one-element-at-a-time sampler that cycle through individual element $d_{j,t}$ over $j = 1, ..., K$ and $t = 1, ..., n$. However such an approach is usually avoided as updating one element a time could lead to sampling inefficiency when K grows. Computational-wise, the cost of the one-element-at-a-time sampler is $\mathcal{O}(Kn)$ and is K times of the $\mathcal{O}(n)$ cost of the proposed approach in this paper.

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 proposed 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 $z_{j,t}$.[2] 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.

[2]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 $z_{j,t}$.

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 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.[3]

2.1 Prior of Parameters

Gamma priors 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 $j = 1, ..., K$.[4] 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

[3] 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.

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

$v_{j,0}^2 = \tau_0\tau_j$ 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 the inverted beta distribution.[5]

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

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, the fully marginalized distribution of d_t conditional on d_{t-1} is not available in closed form. Instead I first examine the distribution of d_t that is conditional on d_{t-1}, μ and ρ but marginalizes over a^2 analytically. By utilizing the prior distribution $a^2 \sim G(0.5, 2\tau_a)$ and the density transformation from z_t to d_t, one can derive:

$$p(d_t|d_{t-1}, \mu, \rho, \tau_a) = \frac{1}{\pi\sqrt{\tau_a}d_t(1-d_t)}\mathcal{K}_0\left(\sqrt{\frac{(\log(d_t) - \log(1-d_t) - \mu_t)^2}{\tau_a}}\right) \tag{5}$$

where $\mathcal{K}_0(\cdot)$ denotes the modified Bessel function of the second kind with the order 0 and $\mu_t = (1-\rho)\mu + \rho\log\left(\frac{d_{t-1}}{1-d_{t-1}}\right)$ denotes the conditional mean of $z_t = \log\left(\frac{d_t}{1-d_t}\right)$. It is illuminating to investigate this partially marginalized distribution of d_t over different values of μ_t and the hyper-parameter τ_a for a^2 in order to understand the shrinkage property of d_t and to guide the choice of the hyper-parameter τ_a.

[5]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 $p(z) = \frac{2}{\pi(1+z^2)}I\{z > 0\}$.

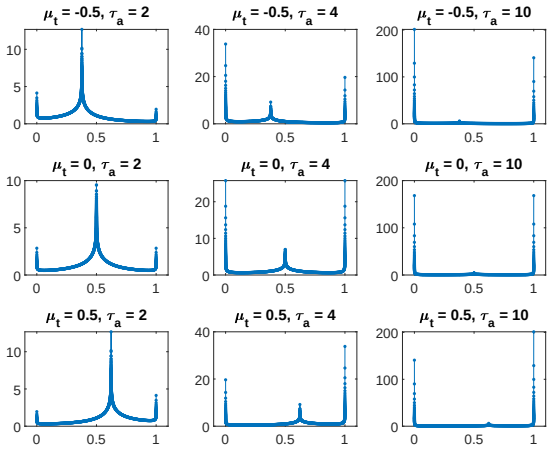
Figure 1 shows the density $p(d_t|d_{t-1}, \mu, \rho, \tau_a)$ over the grid $\mu_t \in \{-0.5, 0, 0.5\}$ and $\tau_a \in \{2, 4, 10\}$. It can be seen that the hyper-parameter τ_a controls the overall shrinkage ability of d_t where a larger τ_a pushes more probability mass of d_t towards the two ends 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 conditional mean μ_t affects the degree of asymmetry in the density of d_t. When $\mu_t = 0$, the conditional density of d_t is symmetric around the central point 0.5. As μ_t 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_t moves towards the positive domain.

Based on the findings in Figure 1, I focus on the case $\tau_a = 10$ and examine the impact of d_{t-1} and ρ on d_t while integrating out the parameter μ numerically based on its distribution specified in Section 2.1. Figure 2 plots the resulting distribution of d_t conditional on the grid of $d_{t-1} \in \{0.1, 0.9\}$ and $\rho \in \{0.1, 0.9\}$. When the serial correlation in the latent variable z_t is weak ($\rho = 0.1$), d_t would essentially switch between the two ends of 0 and 1 with little dependence on d_{t-1}. On the other hand, when $\rho = 0.9$, d_t and d_{t-1} are strongly correlated. Of note, in the case of $\rho = 0.9$ there is still an appreciable probability mass of d_t concentrating on the opposite end of d_{t-1} and hence allowing the behavior of the time-varying coefficients at time t to be different from that at time $t - 1$.

2.3 Closely Related Models

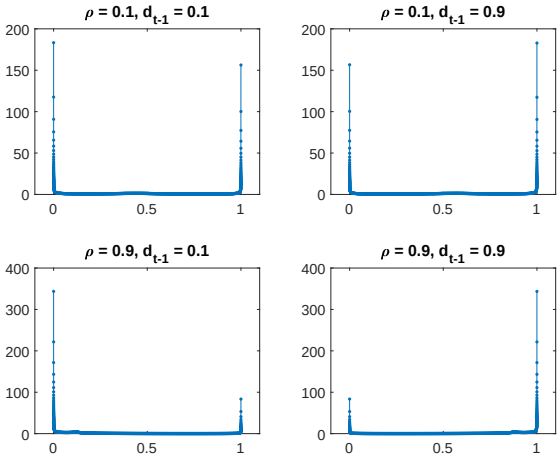
Many previous studies have explored the use of a hard indicator function or a smooth transition function such as the logistic one to model regime changes in regression parameters. Examples in this strand of literature include the threshold regression model of Tong (1990) and the smooth transition regression 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.

Figure 1: Conditional Density of the Mixture Indicator d_t



Note: The figure shows the conditional density $p(d_t|d_{t-1}, \mu, \rho, \tau_a)$ for the mixture indicator $d_t \in (0, 1)$ over different values of the conditional mean parameter $\mu_t = (1-\rho)\mu + \rho \log\left(\frac{d_{t-1}}{1-d_{t-1}}\right)$ and the hyper-parameter τ_a for the conditional variance parameter a^2 where $a^2 \sim G(0.5, 2\tau_a)$.

Figure 2: Serial Dependence of the Mixture Indicator d_t



Note: The figure shows the conditional density $p(d_t|d_{t-1}, \rho, \tau_a = 10)$ for the mixture indicator $d_t \in (0, 1)$ over the grid of $d_{t-1} \in \{0.1, 0.9\}$ and $\rho \in \{0.1, 0.9\}$ that integrates out the parameter a^2 analytically and the parameter μ numerically.

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.[6]

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. 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 (e.g. Hauzenberger et al. (2020), Huber and Pfarrhofer (2021), Pruser (2021)) as a computationally attractive alternative to the spike-and-slab approach in the MI model. Among these studies an 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 (i.e. $d_{j,t} = \exp(z_{j,t})$) and additionally imposing a z distribution for the innovation $\xi_{j,t}$ of each latent variable $z_{j,t}$.[7] Note that the ratio of $\exp(z_{j,t})$ to $\frac{1}{1+\exp(-z_{j,t})}$ equals 1 as $z_{j,t} \to -\infty$. As such, the local component $d_{j,t}$ shrinks the local movement of each coefficient towards zero at the same speed in both the dynamic

[6]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)).

[7]The possibility of specifying $z_{j,t}$ in the LMI model to follow a z distribution, which is a scale mixture of the normal distribution, as in the dynamic horseshoe model is left for future research.

horseshoe and LMI models. On the other hand, as $z_{j,t} \to +\infty$, the local component $d_{j,t}$ is unbounded in the dynamic horseshoe model while approaching 1 in the LMI model, similar to the mixture indicator of a spike-and-slab distribution. In this paper, the forecast performance of the dynamic horseshoe and LMI models 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.[8] 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:

$$\begin{aligned} d_{j,t} &= \frac{1}{1+\exp(-\mu_j - a_j z^*_{j,t})}, \\ z^*_{j,t} &= \rho_j z^*_{j,t-1} + \xi^*_{j,t}, \quad \xi^*_{j,t} \sim N(0,1), \\ z^*_{j,1} &\sim N\left(0, 1/(1-\rho_j^2)\right) \end{aligned} \tag{6}$$

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.

[8]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.

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 $\{\sigma_t^2\}_{t=1}^n$ 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 B. 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 C. In the homoskedastic case $\sigma_t^2 = \sigma^2$ 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).[9] 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:

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

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

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 D.

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.

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 E.

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 $i > i^*$.[10] i^* is a fixed 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:

$$c = \frac{1}{Kp^*(1-p^*)} + \left(1 - \frac{1}{K}\right) \frac{\sqrt{2\pi}\, \exp\left(\frac{\alpha_w^2}{2}\right)}{2\alpha_w} \tag{7}$$

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^*.

[10] 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).

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 F. 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.[11] 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 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 3 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.[12]

To quantify the accuracy of β_t estimates, Figure 4 compares the point-wise root mean squared error (RMSE) of β_t by the MI and LMI models:

$$\text{RMSE}_{j,t} = \sqrt{\frac{1}{M}\sum_{i=1}^{M}\left(\hat{\beta}_{j,t}^{(i)} - \beta_{j,t}\right)^2} \tag{8}$$

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 4 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.

[11] 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.

[12] Appendix A provides the average width of point-wise 90% and 95% credible sets of β_t by the MI and LMI models and shows that they are very close.

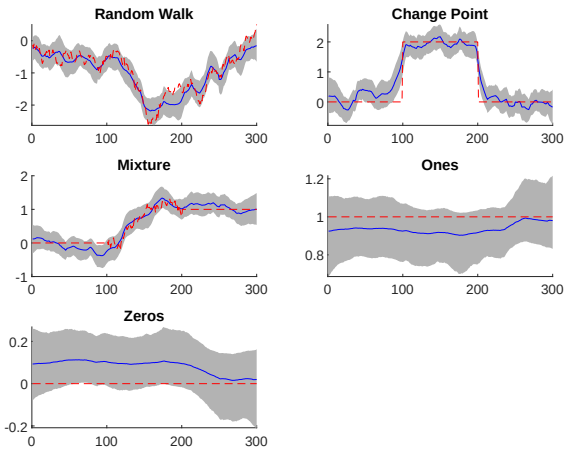
For robustness check, such a comparison of RMSEs of β_t is conducted over 20 data series simulated from the DGP and the results are similar to the ones presented.

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 5 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 6 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 $t = 1, ..., 300$.

Figure 3: Estimate of β_t: Simulation Study



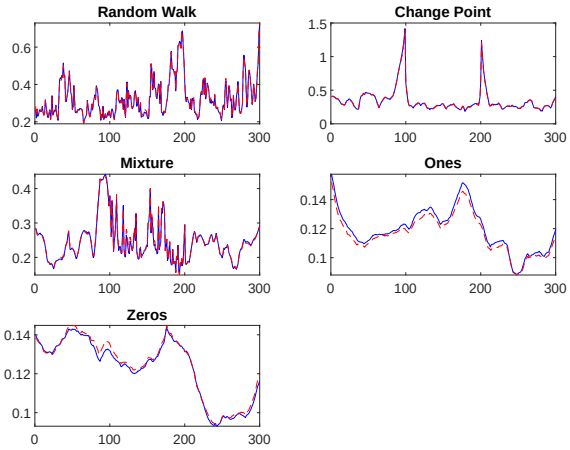
Note: The figure shows the point-wise posterior median (solid line) and 90% credible set (gray area) of estimated coefficients β_t for $t = 1, ..., 300$ by the LMI model. The dashed line is the true coefficient.

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.

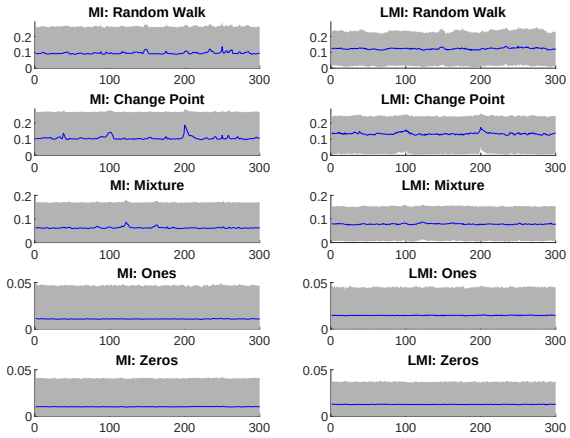
Out-of-sample forecasts are used to test the performance of the LMI model against two alternative TVP models with dynamic shrinkage features. The first alternative model is a restricted version of the MI model (*RMI* hereafter) that places conventional spike-and-slab distributions for the coefficient 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),

Figure 4: Comparing Root Mean Squared Error of β_t Estimates: Simulation Study



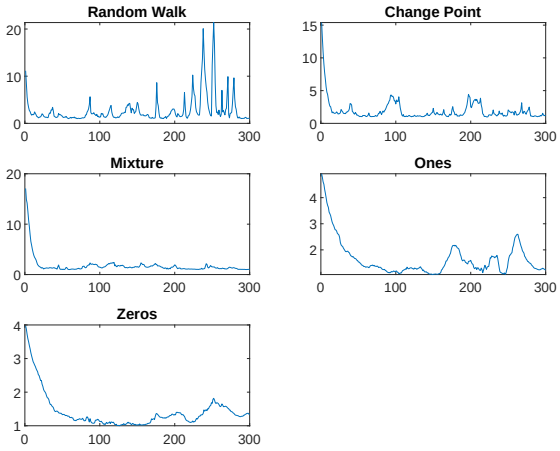
Note: The figure compares the point-wise root mean squared error (Equation (8)) of estimated coefficients β_t for $t = 1, ..., 300$ by the LMI and MI models. The solid line is the root mean squared error by the LMI model while the dashed line is by the MI model.

Figure 5: 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 solid line) and 90% credible set (gray area) of the conditional standard deviations by one of the models under study.

Figure 6: 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.

Huber and Pfarrhofer (2021)). Appendix G 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 Table 1 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.[13] The data sample runs from Q1 1947 to Q4 2020 with a total of 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.

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

Table 1: List of Predictors for Equity Premium

Name	Description
Dividend price ratio	Log dividends minus log price
Dividend payout ratio	Log dividends minus log earnings
Stock variance	Sum of squared daily returns on the S&P500 index
Book-to-market ratio	Ratio of book to market value for the Dow Jones Industrial Average index
Net equity expansion	Ratio of 12-month moving sums of net issues by NYSE listed stocks divided by the total end-of-year market capitalization of NYSE stocks
Treasury bill rate	Quarterly change of 3-month secondary market Treasury bill rate
Long term yield	Quarterly change of long-term government bond yield from Ibbotson's *Stocks, Bonds, Bills and Inflation Yearbook*
Term spread	Long term yield minus treasury bill rate
Default yield spread	Difference between BAA and AAA-rated corporate bond yields
Default return spread	Difference between long-term corporate and government bond returns
Inflation rate	Consumer price index (all urban consumers)
Investment-to-capital ratio	Ratio of aggregate (private non-residential fixed) investment to aggregate capital

Note: The data is publicly available from Amit Goyal's website https://sites.google.com/view/agoyal145/?redirpath=/. Detailed descriptions of the variables can be found in Welch and Goyal (2008).

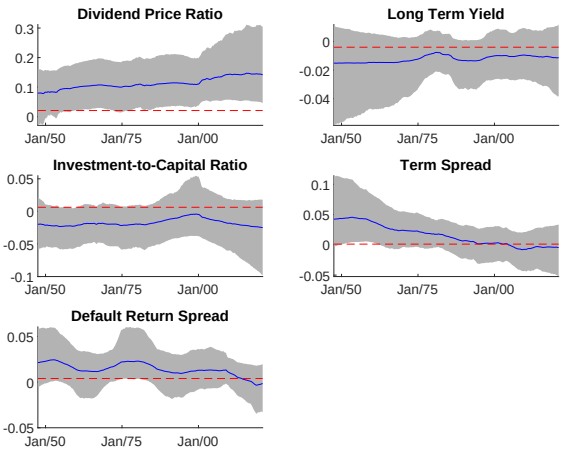
A total of 10,000 posterior draws from the LMI model are collected for analysis after a burn-in length of 2,000. The acceptance rates of MH steps are between 0.23 to 0.26 and are close to the target value of 0.25. Posterior draws of the model parameters mix reasonably well. For example, the inefficiency 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) as the value of zero is consistently outside or near the boundaries of the point-wise 90% credible set of their coefficients. The two predictors *term spread* and *default return spread* were significantly positive at the beginning of the data sample but gradually become insignificant. Figure 7 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

Figure 7: Estimate of Selected β_t: Equity Premium Prediction

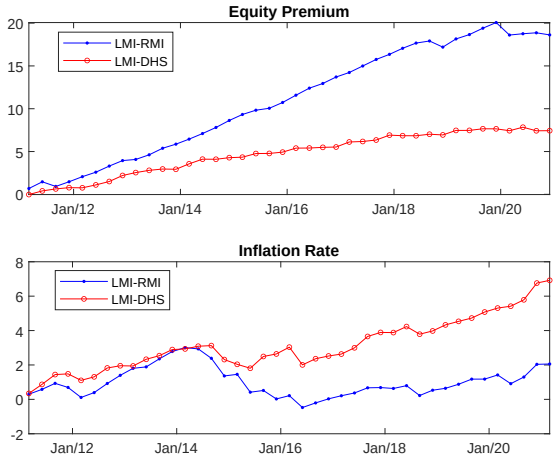


Note: The figure shows the point-wise posterior median (solid line) and 90% credible set (gray area) of estimated coefficient β_t by the LMI model for the 5 regressors in the equity premium prediction model that are either significant throughout the data sample (dividend price ratio, long term yield, 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 Table 1. The dashed line is the static regression estimate for benchmarking.

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).[14] 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.

[14] 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.

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.

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 Table 2. 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.[15] 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

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

Table 2: List of Predictors for Inflation Rate

Name	Description
GDP	Log change of real GDP
Investment	Log change of real gross private domestic investment
Expenditure	Log change of real government consumption expenditures and gross investment
Imports	Log change of imports of goods and services
Potential GDP	Log change of real potential GDP
Employee	Log change of total non-farm employees
Unemployment	Change of unemployment rate
Wage	Log change of average hourly earnings of production and non-supervisory employees
House start	Log change of new privately-owned housing unites started
House supply	Change of the ratio of houses for sale to houses sold
Public debt	Change of the ratio of public debt to GDP
Consumer debt	Log change of consumer credit to households and non-profit organizations
Mortgage	Log change of one-to-four-family residential mortgages
Energy price	Log change of consumer price index for energy in U.S. city average
Producer price	Log change of producer price index for all commodities
Short rate	Change of 3-month Treasury bill rate
Term spread	Difference between 10-year Treasury constant maturity rate and 3-month Treasury bill rate
S&P500	Log change of average daily closing price
M1	Log change of M1 money stock
M2	Log change of M2 money stock

Note: Data on the S&P500 index is from Robert Shiller's website http://www.econ.yale.edu/ shiller/data.htm. Data on all other variables are from the FRED database of the U.S. Federal Reserve Bank of St. Louis.

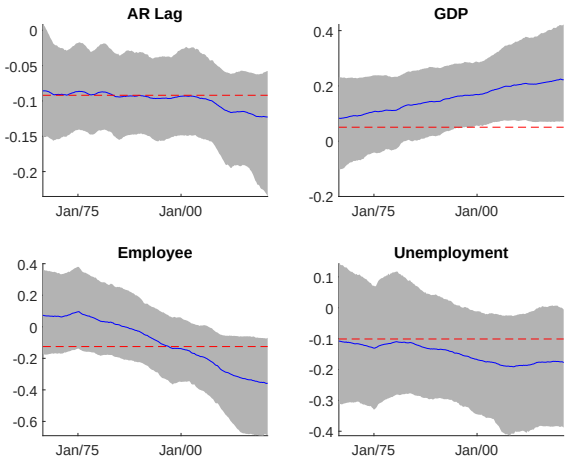
are normalized in estimation.

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 9 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 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 is 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.

Figure 9: 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 Table 2. The dashed line is the static regression estimate for benchmarking.

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 two alternative dynamic shrinkage models: 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.

Appendix

A Additional Results of the Simulation Study

To compare the uncertainty quantification of the MI and LMI models, this section compares the width of the point-wise credible set of the time varying coefficients β_t that is averaged over 20 simulated data samples from the DGP described in Section 4. Beside the 90% credible set, the level of 95% is also examined to compare the behavior of these two models farther into the tail of the posterior distribution of the coefficients. Table A1 summarizes the width of the point-wise credible set of β_t that is averaged over simulations and the time points $t = 1, ..., 300$. It can be seen that the average credible set width of β_t is very close between the MI and LMI models, confirming the findings in Section 4 that the model outputs from the LMI model are comparable to those from the original MI model.

Table A1: Average Width of Point-Wise Credible Set of β_t

	90% Credible Set		95% Credible Set	
	MI	LMI	MI	LMI
Random walk	0.585	0.588	0.758	0.760
Change point	0.648	0.645	0.842	0.840
Mixture	0.441	0.443	0.573	0.575
Ones	0.221	0.221	0.300	0.299
Zeros	0.191	0.192	0.259	0.259

Note: The table compares the average width of point-wise 90% and 95% credible sets of the coefficients β_t estimated by the MI and LMI models. The average width is computed over 20 simulated data sets and the time points $t = 1, ..., 300$.

B Hyper-Parameters of Horseshoe Prior

The horseshoe prior of the initial regression coefficients is $\beta_{j,0} \sim N(0, \tau_0\tau_j)$ with $\tau_0 \sim IB(0.5, 0.5)$ and $\tau_j \sim IB(0.5, 0.5)$ for $j = 1, ..., K$. Following Makalic and Schmidt (2016), the inverted beta distributions are represented as hierarchical inverse gamma ones by in-

troducing auxiliary variables::

$$\tau_0 \sim IB(0.5, 0.5) \Longleftrightarrow \tau_0|\kappa_0 \sim IG\left(0.5, \frac{1}{\kappa_0}\right), \quad \kappa_0 \sim IG(0.5, 1)$$
$$\tau_j \sim IB(0.5, 0.5) \Longleftrightarrow \tau_j|\kappa_j \sim IG\left(0.5, \frac{1}{\kappa_j}\right), \quad \kappa_j \sim IG(0.5, 1)$$

with the following posteriors:

$$\tau_0|\kappa_0, \beta_0, \tau_1, ..., \tau_K \sim IG\left(\frac{1+K}{2}, \frac{1}{\kappa_0} + \frac{1}{2}\sum_{j=1}^{K}\frac{1}{\tau_j}\beta_{j,0}^2\right),$$
$$\kappa_0|\tau_0 \sim IG\left(1, 1 + \frac{1}{\tau_0}\right),$$
$$\tau_j|\beta_0, \tau_0, \kappa_j \sim IG\left(1, \frac{1}{\kappa_j} + \frac{1}{2\tau_0}\beta_{j,0}^2\right),$$
$$\kappa_j|\tau_j \sim IG\left(1, 1 + \frac{1}{\tau_j}\right).$$

C Estimating SV Model

The sampler of Kastner and Fruhwirth-Schnatter (2014) is adapted to estimate the SV model of Equation (3). The log linearization strategy of Omori et al. (2007) is applied to approximate the logarithm of a $\chi^2(1)$-distributed variable by a mixture of normal distributions. A key ingredient of Kastner and Fruhwirth-Schnatter (2014) is applying the ASIS strategy of Yu and Meng (2011) to boost the sampling efficiency of the long-run mean and the variance parameter of the log volatility process. The details of the method can be found in Kastner and Fruhwirth-Schnatter (2014) and are not repeated here to save space.

The main difference in this paper from Kastner and Fruhwirth-Schnatter (2014) is the prior of the variance parameter in the log volatility process. Instead of setting a fixed value for the scale parameter s_h in the gamma prior $\sigma_h^2 \sim G(0.5, 2s_h)$, this paper specifies a prior $s_h \sim IB(0.5, 0.5)$ to determine s_h in a data driven way. The conditional posterior of s_h can be obtained by applying the hierarchical inverse gamma representation in Makalic and Schmidt (2016): $s_h|a_h, \sigma_h^2 \sim IG\left(1, \frac{1}{a_h} + \frac{\sigma_h^2}{2}\right)$ where a_h is an auxiliary variable with the prior $a_h \sim IG(0.5, 1)$ and the posterior $a_h|s_h \sim IG\left(1, 1 + \frac{1}{s_h}\right)$.

D Sampling the Latent Variable z_t^*

The target is to sample from the posterior $p(z_t^*|y, x, \Theta, z_{-t}^*) \propto p(z_t^*|z_{-t}^*, \rho)p(y|x, z^*, \Theta)$ following the notations from Section 3.1.

Given the AR specification of z_t^* in Equation (6), the prior part can be written as $p(z_t^*|z_{-t}^*, \rho) \propto p(z_t^*|z_{t-1}^*, \rho)p(z_{t+1}^*|z_t^*, \rho)$ for $t = 2, ..., n-1$. One can derive $z_t^*|z_{-t}^*, \rho \sim N(b_{z,t}, B_{z,t})$ where $B_{z,t}$ is a K-by-K diagonal matrix with the j^{th} diagonal entry $\frac{1}{1+\rho_j^2}$ and $b_{z,t}$ is a K-by-1 vector with the j^{th} entry $\frac{\rho_j}{1+\rho_j^2}(z_{j,t-1} + z_{j,t+1})$ for $j = 1, ..., K$ and $t = 2, ..., n-1$. When $t = 1$, one has $p(z_1^*|z_{-1}^*) \propto p(z_1^*)p(z_2^*|z_1^*)$ where $z_{j,1}^* \sim N\left(0, \frac{1}{1-\rho_j^2}\right)$ for $j = 1, ..., K$. It is straightforward to derive $z_1^*|z_{-1}^* \sim N(b_{z,1}, B_{z,1})$ with $B_{z,1} = I_K$ and $b_{z,1} = \rho \odot z_2^*$, where $\odot$ denotes the Hadamard product. When $t = n$, the posterior is simply $p(z_n^*|z_{-n}^*) \propto p(z_n^*|z_{n-1}^*)$ which is a normal distribution $N(\rho \odot z_{n-1}^*, I_K)$.

Denote $y^t = \{y_1, ..., y_t\}$ and $y^{t,n} = \{y_t, ..., y_n\}$. The GCK algorithm is applied to compute the components of the likelihood $p(y|x, z^*, \Theta)$ that is relevant to sampling z_t^*:

$$\begin{aligned} p(y|x, z^*, \Theta) &\propto p(y^{t,n}|y^{t-1}, x, z^*, \Theta) \\ &\propto r_t^{-\frac{1}{2}} \det(Q_t)^{-\frac{1}{2}} \exp\left(-\frac{1}{2} m_t' \Omega_t m_t + \mu_t' m_t + \frac{1}{2}\phi_t' Q_t^{-1}\phi_t - \frac{(y_t - v_t)^2}{2r_t}\right) \end{aligned} \quad \text{(D1)}$$

where $m_t = E(\beta_t|y^t, x, z^*, \Theta)$ and $M_t = V(\beta_t|y^t, x, z^*, \Theta)$ are computed by a Kalman filter, $r_t = q_t + x_t' M_{t-1} x_t$, $q_t = \sigma_t^2 + x_t' W_t x_t$, $W_t = \text{diag}(d_t \odot v^2)$, $Q_t = I_K + T_t' \Omega_t T_t$, $T_t T_t' = M_t$, $\phi_t = T_t'(\mu_t - \Omega_t m_t)$ and $v_t = x_t' m_{t-1}$. The quantities of μ_t and Ω_t are computed by a backward recursion:

$$\begin{aligned} &\Omega_n = 0, \;\; \mu_n = 0 \\ &\Omega_{t-1} = A_t'(\Omega_t - \Omega_t C_t D_t^{-1} C_t' \Omega_t) A_t + \frac{x_t x_t'}{q_t} \\ &\mu_{t-1} = A_t'(I_K - \Omega_t C_t D_t^{-1} C_t')(\mu_t - \Omega_t b_t y_t) + \frac{x_t y_t}{q_t} \end{aligned} \quad \text{(D2)}$$

where $b_t = \frac{W_t x_t}{q_t}$, $A_t = I_K - b_t x_t'$, $C_t' C_t = W_t - \frac{W_t x_t x_t' W_t}{q_t}$ and $D_t = I_K + C_t' \Omega_t C_t$. The details of derivation of the GCK algorithm can be found in the original paper of Gerlach et al. (2000) and are not repeated here to save space. The Kalman filter is a standard technique for linear Gaussian state space systems for which a concise description can be found in Gerlach et al. (2000) and a textbook treatment can be found in Hamilton (1994).

E Computing the Model Likelihood Function

The likelihood function $p(y|x, z^*, \Theta)$ of the LMI model (Equation (2) and (4)) can be computed by running a Kalman filter to integrate out β_t. Specifically, decompose the likelihood as $p(y|x, z^*, \Theta) = p(y_1|x, z^*, \Theta) \prod_{t=1}^{n-1} p(y_{t+1}|y^t, x, z^*, \Theta)$ where $y^t = \{y_1, ..., y_t\}$.

To compute $p(y_{t+1}|y^t, x, z^*, \Theta)$, first consider the distribution $p(y_{t+1}|\beta_t, y^t, x, z^*, \Theta)$. By substituting $\beta_{t+1} = \beta_t + \eta_{t+1}$ into the equation $y_{t+1} = x'_{t+1}\beta_{t+1} + \epsilon_{t+1}$, it is straightforward to show $y_{t+1}|\beta_t, y^t, x, z^*, \Theta \sim N(x'_{t+1}\beta_t, \sigma^2_{t+1} + x'_{t+1}W_{t+1}x_{t+1})$ where $W_{t+1} = \text{diag}(d_{t+1} \odot v^2)$. Next write $p(y_{t+1}|y^t, x, z^*, \Theta) = \int p(y_{t+1}|\beta_t, y^t, x, z^*, \Theta)p(\beta_t|y^t, x, z^*, \Theta)d\beta_t$. The distribution $p(\beta_t|y^t, x, z^*, \theta_z)$ is normal where the mean $m_t = E(\beta_t|y^t, x, z^*, \Theta)$ and the covariance matrix $M_t = V(\beta_t|y^t, x, z^*, \Theta)$ can be computed through a Kalman filter. It can be shown that $p(y_{t+1}|y^t, x, z^*, \Theta)$ is also normal with the mean $x'_{t+1}m_t$ and the variance $x'_{t+1}M_t x_{t+1} + \sigma^2_{t+1} + x'_{t+1}W_{t+1}x_{t+1}$.

The initial component $p(y_1|x, z^*, \Theta) = \int p(y_1|\beta_0, x, z^*, \Theta)p(\beta_0|x, z^*, \Theta)d\beta_0$ is computed by inserting the prior $p(\beta_0|v_0^2) = N(0, \text{diag}(v_0^2))$, where $v_0^2 = \tau_0\tau$, and is normal with the mean of zero and the variance $x'_1\text{diag}(v_0^2)x_1 + \sigma_1^2 + x'_1W_1x_1$.

F ASIS for the LMI Model

The ASIS is applied to two blocks of parameters. The first block involves the time-varying coefficients β and the parameters v and β_0 and is based on the following reparameterization of Equation (2):

$$\begin{aligned} y_t &= x'_t\beta_0 + (x_t \odot \beta_t^*)'v + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma_t^2), \\ \beta_t^* &= \beta_{t-1}^* + \eta_t^*, \quad \eta_t^* \sim N(0, \text{diag}(d_t)), \quad \beta_0^* = 0 \end{aligned} \tag{F1}$$

where $\beta_t^* = \text{diag}(v)^{-1}(\beta_t - \beta_0)$. Note that β_0 and v become the fixed coefficients of a linear regression model conditional on β_t^* and σ_t^2 in Equation (F1). Such a reparameterization of a TVP model was pioneered in Fruhwirth-Schnatter and Wagner (2010). The specific steps are as follows:

1. In each MCMC sweep after running the Gibbs sampler of Section 3.1, compute $\beta_t^* = \text{diag}(v)^{-1}(\beta_t - \beta_0)$ for $t = 1, ..., n$.

2. Let α be a $2K$-by-1 vector stacking β_0 and v. Conditional on β_t^*, σ_t^2 and the hyper-parameters θ_0, draw α from a linear regression with the posterior $N(b_\alpha, B_\alpha)$, where $B_\alpha^{-1} = B_0^{-1} + \sum_{t=1}^n \frac{1}{\sigma_t^2}\tilde{x}_t\tilde{x}_t'$, $B_\alpha^{-1}b_\alpha = \sum_{t=1}^n \frac{1}{\sigma_t^2}\tilde{x}_t y_t$, B_0 is a $2K$-by-$2K$ diagonal matrix with the diagonal elements $\tau_0\tau_1$, ..., $\tau_0\tau_K$, τ_v, ..., τ_v, and $\tilde{x}_t = [x_t \;\; x_t \odot \beta_t^*]'$.

3. Compute back $\beta_t = \beta_t^* \odot v + \beta_0$ for $t = 1, ..., n$.

The resulting β_0, v and β_t from this ASIS step are used as their final draw in an MCMC sweep.

The second ASIS boosting is applied to the block of the latent variable z^* and the parameters μ and a and is based on the AR process for the original latent variable $z = \{z_t\}_{t=1}^n$:

$$\begin{aligned} z_t &= (\mathbf{1}_K - \rho) \odot \mu + \rho \odot z_{t-1} + \xi_t, \quad \xi_t \sim N(0, \text{diag}(a^2)), \\ z_1 &\sim N(\mu, (I_K - \text{diag}(\rho^2))^{-1}\text{diag}(a^2)) \end{aligned} \tag{F2}$$

where $\mathbf{1}_K$ denotes a K-by-1 vector of ones. The steps are as follows:

1. In each MCMC sweep after running the Gibbs sampler of Section 3.1 and the first ASIS boosting, compute $z_t = \mu + a \odot z_t^*$ for $t = 1, ..., n$.

2. Draw μ from its posterior $p(\mu|z, \rho, a, \psi)$ based on Equation (F2) and the prior $\mu \sim N(0, \text{diag}(\psi))$. The resulting posterior is a normal distribution $N(B_\mu^{-1}b_\mu, B_\mu^{-1})$ where the j^{th} entry of b_μ is $\frac{1}{a_j^2}\left((1-\rho_j^2)z_{j,1} + (1-\rho_j)\sum_{t=1}^{n-1}(z_{j,t+1} - \rho_j z_{j,t})\right)$ and B_μ is a diagonal matrix with the j^{th} diagonal entry $\frac{1}{\psi_j} + \frac{1}{a_j^2}\left(1 - \rho_j^2 + (n-1)(1-\rho_j)^2\right)$ for $j = 1, ..., K$.

3. Keep the sign of a.

4. Draw a^2 from its posterior $p(a^2|z, \rho, \mu)$ based on Equation (F2) and the prior $a_j^2 \sim G(0.5, 2\tau_a)$ for $j = 1, ..., K$. It can be derived that the posterior for each a_j^2 is a generalized inverse Gaussian distribution $GIG(\frac{1-n}{2}, \frac{1}{\tau_a}, \sum_{t=1}^n \xi_{j,t}^2)$ where $\xi_{j,1} = \sqrt{1-\rho_j^2}(z_{j,1} - \mu_1)$ and $\xi_{j,t} = z_{j,t} - (1-\rho_j)\mu_j - \rho_j z_{j,t-1}$ for $t = 2, ..., n$ and $j = 1, ..., K$. Update a as the square root of a^2 times the sign from the previous step.

5. Compute back $z_t^* = \text{diag}(a)^{-1}(z_t - \mu)$ for $t = 1, ..., n$.

The resulting μ, a and z_t^* from this ASIS step are used as their final draw in an MCMC sweep.

G Alternative TVP Models with Dynamic Shrinkage

Two alternative TVP models with dynamic shrinkage features are compared to the proposed LMI model in the applications of this paper.

The RMI model is a restricted version of the MI model of Giordani and Kohn (2008) which is specified in Equation (2). Instead of the unrestricted 2^K scenarios in a model with K regressors, the RMI model allows only 3 scenarios for the mixture indicator $d_t = [d_{1,t} \ ... \ d_{K,t}]'$ at each time t: all $d_{j,t}$ equal zero (i.e. no parameter change at time t), all $d_{j,t}$ equal one (i.e. all parameters change at time t), and $d_{1,t} = 1$ and $d_{j,t} = 0$ for $j > 1$ (i.e. only the intercept changes at time t while other parameters remain constant). Denote the 3 scenarios as s_1, s_2 and s_3. The dynamics of d_t is Markovian following $p(d_t = s_j | d_{t-1} = s_j) = p_r$ and $p(d_t = s_i | d_{t-1} = s_j) = \frac{1-p_r}{2}$ for $i \neq j$. The prior for p_r is Beta$(50, 0.5)$ to favor persistent dynamics of d_t. Priors for other model parameters are the same as those in the LMI model.

The DHS model of Kowal et al. (2019) follows:

$$
\begin{aligned}
& y_t = x_t' \beta_t + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma_t^2), \\
& \Delta \beta_{j,t} \sim N(0, \tau_{v,0} \tau_{v,j} \phi_{j,t}), \quad \beta_{j,0} \sim N(0, v_{j,0}^2), \\
& \log(\phi_{j,t}) = \rho_j \log(\phi_{j,t-1}) + \xi_{j,t}, \quad \xi_{j,t} \sim Z(0.5, 0.5, 0, 1), \quad \phi_{j,0} = 1
\end{aligned} \tag{G1}
$$

where $\sqrt{\tau_{v,0}} \sim C^+(0, \frac{1}{\sqrt{nK}})$ and $\sqrt{\tau_{v,j}} \sim C^+(0, 1)$ for $j = 1, ..., K$. The logarithm of the local variance $\phi_{j,t}$ follows an autoregressive process with a z-distributed innovation $\xi_{j,t}$ which is obtained as the logarithm of an inverted-beta distributed random variable. The z distribution can be sampled as a scale mixture of normal distributions. See Kowal et al. (2019) for the motivation and details of the DHS model. In the applications, the prior for ρ_j is $N(0.95, 1) I\{-1 < \rho_j < 1\}$. The priors of other model parameters are the same as those in the LMI model.

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., 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 36*, 703–727.

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.

Giordani, P., R. Kohn, and D. van Dijk (2007). A unified approach to nonlinearity, structural change, and outliers. *Journal of Econometrics 137*, 112–133.

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.

Omori, Y., S. Chib, N. Shephard, and J. Nakajima (2007). Stochastic volatility with leverage: Fast and efficient likelihood inference. *Journal of Econometrics 140*, 425–449.

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

Pruser, J. (2021). The horseshoe prior for time-varying parameter VARs and monetary policy. *Journal of Economic Dynamics and Control 129*, 104188.

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.

Conflict of Interest Statement

No conflict of interests