Time Dependent Shrinkage of Time-Varying Parameter Regression Models

Article by Zhongfang He

July 4, 2022

Time-varying vector auotregressions are arguably the most successful highly-parameterized,nonlinear models used in macroeconomics and related disciplines today. As is common in the machine-learning literature these methods start by specifying more parameters than one can realistically need and then adding regularization to reduce the effective degrees of freedom. Unlike many models developed in industry and the computer science literature, this literature focuses on understanding precisely want data generating processes (DGPs) that can be estimated.

This paper studies DGPS of the following form:

$$\begin{aligned} y_t &= x_t'\beta_t + \sigma_t N(0,1) \\ \beta_t &= \beta_{t-1} + \text{diag}w_t{}^{1/2}N(0,I) \\ \beta_0 &\sim N(0, \text{diag}w_0) \\ w_t &\sim G \end{aligned} \tag{1}$$

This paper focuses what is an appropriate way to specify G. The first leading special case is $w_t = 0$; that is, we have a constant coefficients model (we're just doing linear regression). The second leading special case is $w_t = c > 0$. This is the now standard time-varying parameter model with a random walk coefficients. The third leading special case is $w_t = c > 0 for in \mathcal{T}$ and $w_t = 0$ otherwise for some set $\mathcal{T}$.

The main contribution of this paper is developing a prior that lets the data endogenously determine how w_t behaves. This problem is called the *variance selection* problem. The most successful shrinkage and selection method is the LASSO. The Bayesian LASSO places zero prior weight on the coefficient being zero. So Carvalho, Polson, and Scott (2009) proposed the horseshoe prior to place positive mass on 0 while still allowing for shrinkage for non-zero parameters. Various authoors have extended this idea to the variance selection problem. Kowal, Matteson, and Ruppert (2019) proposed the dynamic horseshoe prior to allow for non-trivial volatility clustering in w_t over time. This paper develops a new prior—the gamma horseshoe prior—that places even more weight on zero and has even fatter tails than the horseshoe prior and develops a computationally efficient estimator for it.

In particular, this paper takes the horseshoe prior:

$$w_{j,t} \sim G(w_{j,t}; 0.5, 2v_j d_{j,t}) IB(d_{j,t}; 0.5, 0.5) G(v_j; 0.5, 2\tau_0\tau_j) IB(\tau_0; 0.5, 0.5) IB(\tau_j; 0.5, 0.5), \tag{2}$$

where IB is the inverse-beta distribution and G is the gamma distribution. In contrast, in the standard horseshoe prior

$$w_{j,t} \sim \delta(w_{j,t} = d_{j,t}v_j) IB(d_{j,t}; .5, .5) IB(v_j; .5, .5) \tag{3}$$

That is, the this paper replaces the Dirichlet delta relating $w_{j,t}$to $d_{j,t}v_j$ with a Gamma distribution (i.e., it introduces another shock). It keeps the distribution for the the local shrinkage parameters $d_{j,t}$ the same $IB(d_{j,t}; .5, .5)$ It replaces the distribution for the global shrinkage parameter v_j with the convolution of a Gamma distribution with two inverse beta distributions: $IB(v_j; .5, .5) \rightarrow G(v_j; 0.5, 2\tau_0\tau_j) IB(\tau_0; 0.5, 0.5) IB(\tau_j; 0.5, 0.5)$.

1 estimation

In principle, one could use a standard Gibbs sampler to estimate the proposed model above. However, it is not that surprising the implied sampler exhibits unsatisfactory mixing performance. So the author develops an ancillary-sufficiency interweaving

strategy (ASIS) to sample from the posterior, adapting Yu and Meng 2011.

2 Comments

2.1 Understanding the relationship between various parameterizations

It is quite confusing what the precise relationship between the Horseshoe and Gamma horseshoe prior actually is. For example, in Table 1, the author describes the horseshoe prior as IB(0.5, 0.5) IB(0.5, 05.) IB(0.5, 0.5). However, if you look at Carvalho, Polson, and Scott 2009, a horseshoe prior is defined as IB(0.5, 0.5) IB(0.5, 0.5). A product of three random variables does not have the same distribution as a product of two. This confusion may be in the literature already. The literature appears to be abusing notation. However, that increases (not decreases) the importance of being very careful to explain the results / assumptions here and how they relate to the previous literature.

Furthermore, the comparison of the various parameterizations, (i.e., what precisely is being assumed) is placed in Appendix A and Table 1 after the references. This is arguably the most important part of the paper; it is very difficult to judge the usefulness of the approach or the marginal contribution if that if you do not know what precisely is being proposed. This paper would be much easier to understand if the material in Appendix A and Table 1 was moved into the state variances section.

2.2 Simulation and Empirical results

The simulation section appears fairly well done. I liked that they included both constant, random walk, and various changepoint simulation data generating processes. The empirical sections are fine. The proposed methods do seem to be performing fairly well relative to the horseshoe and dynamic horseshoe priors. A little bit more work into what parts of the distribution are being fit well would be interesting. The author implies that it is coming from more accurate estimation of the β_t. That is

not obvious in general. They may be capturing tail risk instead. It would be interesting to see some metrics such as mean absolute error (MAE) and mean square error (MSE) for the point forecasts. If they are doing a better job at estimating the coefficients instead of modeling the distribution of the coefficients that would be useful. Either answer could be interesting by helping practitioners understand when this model is a good choice.

2.3 Additional theoretical questions

More generally, the coefficients β_t are not "true" underlying objects. The real world is not a time-varying parameter model. They are changing the time-varying tail dependence for y_t. It would be interesting to see what precise classes of DGPs for y_t they can approximate well that other models could not. This is, realistically, quite difficult, but it would be an interesting avenue for future research.

2.4 Time-varying Vector Autogressions TVP-VARs

They mention at the end that this methods could be extended to the multivariate case. This would be quite useful because VARs are arguably the most important case for these sorts of models. I would expect far more people to use this method in practice if this was done.

References

Carvalho, Carlos M, Nicholas G Polson, and James G Scott. 2009. "Handling sparsity via the horseshoe". In *Artificial Intelligence and Statistics*, 73–80. Proceedings of Machine Learning Research.

Kowal, Daniel R., David S Matteson, and David Ruppert. 2019. "Dynamic shrinkage processes". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 81 (4): 781–804.

Yu, Yaming, and Xiao-Li 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.