



Recent Advances in Dynamic Shrinkage in High-Dimensional Regression Models

Authors: GUILHERME COLOMBO SOARES 
– Department of Economics, FEARP, University of São Paulo,
Brazil
guilhermecsoares@usp.br

MÁRCIO POLETTI LAURINI  
– Department of Economics, FEARP, University of São Paulo,
Brazil
laurini@fearp.usp.br

Received: Month 0000

Revised: Month 0000

Accepted: Month 0000

Abstract:

- This survey reviews recent advances in dynamic shrinkage models for time-varying parameter estimation and variable selection in time series. We introduce classical shrinkage methods, including ridge, lasso, elastic net, Bayesian lasso, horseshoe, and spike-and-slab priors, and emphasize their dynamic extensions. These approaches allow shrinkage to evolve over time, improving adaptability to structural breaks and regime changes. We discuss Bayesian and frequentist formulations, computational strategies, and empirical applications in modern econometrics.

Keywords:

- *dynamic shrinkage; time-varying parameters; variable selection; time series; bayesian shrinkage; econometrics.*

AMS Subject Classification:

- 62M10, 62J07, 62F15, 62P20

1. INTRODUCTION

The increasing availability of high-dimensional datasets in macroeconomics and finance, combined with restricted computational capabilities, has posed a central methodological challenge: how can we effectively select relevant variables in dynamic environments? While traditional variable selection techniques have played a pivotal role in reducing model complexity and improving forecasting performance, their static nature often limits their ability to adapt to time-varying data structures and changing economic regimes.

Macroeconomic and financial datasets are particularly prone to structural changes and exhibit features such as parameter instability, regime shifts, and varying signal strength across time. These characteristics motivate the need for shrinkage methods that not only regularize models but also adapt their structure dynamically. Classical shrinkage techniques, such as Ridge (Hoerl and Kennard, 2000), Lasso (Tibshirani, 1996), Elastic Net (Zou and Hastie, 2005) and Bayesian structures like Spike-and-Slab priors (George and McCulloch, 1993) or horseshoe priors (Carvalho et al., 2010) have laid the foundation for controlling overfitting and enhancing interpretability by shrinking or excluding irrelevant variables. However, these approaches are inherently static and do not account for the temporal evolution of variable relevance.

The literature has progressively advanced toward dynamic shrinkage models that allow for time-varying sparsity, where coefficients can enter and exit the model over time depending on their predictive contribution. This extension is particularly relevant in forecasting applications where the set of relevant predictors is not constant across different periods, as demonstrated in seminal contributions such as Stock and Watson (2007) and Cogley and Sargent (2005). These works show that economic relationships evolve over time, and fixed-parameter models fail to capture the dynamic nature of macroeconomic systems.

Recent developments in Bayesian econometrics have introduced flexible frameworks that embed shrinkage priors into state-space models, enabling dynamic variable selection and temporal shrinkage. Methods such as the Dynamic Spike-and-Slab (DSS) priors (Ročková and McAlinn, 2021; Bai et al., 2021; Koop and Korobilis, 2023) or dynamic horseshoe priors (Hauzenberger et al., 2024) integrate hierarchical priors and latent indicators to regulate the inclusion of variables over time. These models are capable of inducing vertical sparsity (activation or deactivation of coefficients across time) and horizontal shrinkage (global regularization across variables), offering a principled approach managing large dimensional time series models under structural change.

This paper is primarily intended as a methodological survey of recent developments in dynamic shrinkage for time-varying parameter regression models. Our goal is to synthesize the rapidly growing literature on dynamic sparsity priors, discuss the main computational strategies used for their estimation, and highlight their relative advantages and limitations. The numerical examples included in the paper are meant to illustrate the qualitative behavior of these methods rather than to provide a comprehensive empirical comparison.

The remainder of this paper is structured as follows. Section 2 reviews classical static shrinkage methods, covering both their traditional formulations and Bayesian interpreta-

tions. Section 3 extends the discussion to dynamic frameworks, with a focus on state-space representations and dynamic sparsity priors. This section is divided into four parts: Part 1 presents the Expectation-Maximization algorithm of [Ročková and McAlinn \(2021\)](#); Part 2 introduces the Variational Bayes model proposed by [Koop and Korobilis \(2023\)](#); Part 3 discusses a more advanced formulation of Variational Bayes developed by [Bernardi et al. \(2023\)](#); and Part 4 compares the forecasting performance of these models. Section 4 presents the scalable MCMC algorithm proposed by [Hauzenberger et al. \(2024\)](#). Finally, Section 5 concludes with a critical assessment of current limitations and outlines directions for future research.

2. STATIC SHRINKAGE

The concept of shrinkage in statistics traces its formal origins to the groundbreaking work of James and Stein in 1961 ([Stein, 1956](#); [James and Stein, 1961](#)). In their seminal paper, they demonstrated that, under a quadratic loss function, the traditional maximum likelihood estimator for the mean of a multivariate normal distribution is inadmissible when the dimension is three or greater. They proposed a class of estimators, now known as James-Stein estimators, which "shrink" ([James and Stein, 1961](#)) the sample mean toward a fixed point, typically the zero, resulting in uniformly lower expected squared error. This counterintuitive result challenged conventional statistical thinking and laid the theoretical foundation for a wide range of shrinkage and regularization techniques developed in subsequent decades.

Pioneering works such as those by [George and McCulloch \(1993\)](#), [Tibshirani \(1996\)](#), and [Hoerl and Kennard \(2000\)](#) introduced methods like Spike-and-Slab, Lasso, and Ridge for variable selection. Today, these methods are well known, extensively discussed in the literature, and widely used in both academic and non-academic applications.

In this section, we provide a brief introduction to both frequentist and Bayesian approaches to static shrinkage. These introductions will serve as a foundation for understanding the dynamic variants discussed in the next section.

We can generalize shrinkage solutions as follows. Suppose a linear model of the form $y = X\beta + \varepsilon$. In problems where the number of covariates exceeds the number of observations, the covariance matrix becomes singular and the OLS estimator becomes non-identifiable, and we cannot obtain a solution using the standard linear model. To address such cases, we introduce a penalization structure into the parameter estimation:

$$(2.1) \quad \hat{\beta} = \arg \min_{\beta} \left\{ \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^p |\beta_j|^p \right\}.$$

As shown by [Hastie et al. \(2009\)](#), this formulation essentially reduces to choosing the type of ℓ_p penalization applied to the coefficients.

2.1. Ridge

In 1970, [Hoerl and Kennard \(2000\)](#) introduced the Ridge Regression, a quadratic type of penalization, which allows estimation of ill-conditioned or nearly singular models. Basically, the Ridge regression aims to solve the problem,

$$Y = \beta X + \epsilon$$

we aim to estimate β as

$$(2.2) \quad \hat{\beta}_{\text{ridge}} = \arg \min_{\beta} \left\{ \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^p \beta_j^2 \right\}$$

where the penalty follows a ℓ_2 -norm structure,

$$(2.3) \quad \text{pen}_{\text{ridge}}(\beta) = \lambda \sum_{j=1}^p \beta_j^2,$$

$\lambda \geq 0$ is a parameter that controls the amount of shrinkage applied to the coefficients. In the limit as $\lambda \rightarrow 0$, we recover the standard linear model, and as $\lambda \rightarrow \infty$, all coefficients are shrunk toward zero. However, with ℓ_p penalties where $p > 1$ —as is the case with Ridge regression—the function is differentiable at zero, and the coefficients β_j are not shrunk exactly to zero. Ridge regression can be interpreted as a *maximum a posteriori* (MAP) estimate in a Bayesian linear regression framework. In this setting, the regression coefficients β are assumed to have independent Gaussian priors:

$$\beta_j \sim \mathcal{N}(0, \tau^2).$$

This prior leads to a log-posterior that combines the usual squared error loss with a penalty term proportional to $\|\beta\|^2$. Maximizing the posterior is therefore equivalent to minimizing:

$$\|y - X\beta\|^2 + \lambda \|\beta\|^2$$

where $\lambda = \sigma^2/\tau^2$, linking the ridge penalty to the ratio of noise variance to prior variance. Thus, ridge regression reflects the prior belief that the coefficients should be small, effectively shrinking them toward zero.

As presented by [Hastie et al. \(2009\)](#), the Ridge model is not equivariant under scaling of the inputs, so it is common practice to standardize the covariates before estimating the model.

2.2. Lasso

The Least Absolute Shrinkage and Selection Operator (Lasso), introduced by [Tibshirani \(1996\)](#), is a penalized regression method closely related to Ridge regression. However, instead of using an ℓ_2 -norm penalty, Lasso employs an ℓ_1 -norm penalty, which allows for variable selection by shrinking some coefficients exactly to zero.

$$(2.4) \quad \hat{\beta}_{\text{lasso}} = \arg \min_{\beta} \left\{ \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^p |\beta_j| \right\},$$

where the penalty follows,

$$(2.5) \quad \text{pen}_{\text{Lasso}}(\beta) = \lambda \sum_{j=1}^p |\beta_j|.$$

This ℓ_1 type penalty make the solutions non-linear in y_i and then there is no closed-form solution and requires numerical optimizations methods to solve (Hastie et al., 2009).

2.3. Elastic-Net

Zou and Hastie (2005) proposed a penalization approach that combines the strengths of both Ridge and Lasso methods. The goal is to merge the computational stability and solution tractability of Ridge regression with the variable selection property of Lasso, which can shrink coefficients exactly to zero. The resulting Elastic Net penalty is a convex combination of the ℓ_1 and ℓ_2 norms:

$$(2.6) \quad \text{Pen}_{\text{Elastic-Net}} = \lambda \sum_{j=1}^p \left(\alpha \beta_j^2 + (1 - \alpha) |\beta_j| \right).$$

It is straightforward to observe that when $\alpha = 0$, the Elastic Net reduces to the Lasso penalty, and when $\alpha = 1$, it becomes equivalent to the Ridge penalty. As noted by Hastie et al. (2009), the Elastic Net retains the ability of Lasso to perform variable selection, while also encouraging grouped shrinkage of correlated predictors, a property characteristic of Ridge regression.

2.4. Bayesian priors

In Bayesian statistics, shrinkage is typically achieved through the specification of informative priors within a hierarchical model structure. By imposing prior beliefs that favor small or zero coefficients, the model can effectively regularize parameter estimates and induce sparsity. This approach is particularly powerful in large or high-dimensional settings, where the number of predictors may exceed the number of observations.

Hierarchical priors allow for flexible modeling of uncertainty by introducing global and local shrinkage parameters. Global shrinkage parameters control the overall level of regularization across all coefficients, while local parameters allow individual coefficients to deviate from the global trend when supported by the data. This framework underlies several popular Bayesian shrinkage priors, such as the Bayesian Lasso, Horseshoe, and Spike-and-Slab, each offering distinct trade-offs between sparsity, bias, and computational complexity.

The choice of prior plays a fundamental role in shaping the behavior of the posterior distribution and, consequently, the model's ability to perform variable selection and prediction. In the following subsections, we present and compare several Bayesian priors designed to achieve shrinkage under different structural assumptions.

2.4.1. Bayesian lasso and ridge

The Bayesian Lasso, proposed by [Park and Casella \(2008\)](#), can be viewed as a Bayesian formulation of the classical Lasso estimator introduced by [Tibshirani \(1996\)](#). As pointed out by both [Park and Casella \(2008\)](#) and [Tibshirani \(1996\)](#), Lasso estimates can be interpreted as posterior mode estimates when the regression coefficients are assigned independent Laplace (double-exponential) priors.

This formulation can be equivalently expressed using a hierarchical representation as a scale mixture of normal distributions, where the Laplace prior arises from integrating out an exponential mixing distribution over the local variances. The full Bayesian hierarchical model is:

$$(2.7) \quad \begin{aligned} \beta_j \mid \tau_j^2, \sigma^2 &\sim \mathcal{N}(0, \sigma^2 \tau_j^2), \\ \tau_j^2 \mid \lambda^2 &\sim \text{Exponential}\left(\frac{\lambda^2}{2}\right), \quad j = 1, \dots, p, \\ \lambda &\sim \text{half-Cauchy}(0, 1). \end{aligned}$$

By integrating out the local variance components τ_j^2 , we recover the marginal Laplace distribution for β_j :

$$(2.8) \quad \beta_j \mid \lambda, \sigma \sim \text{Laplace}\left(0, \frac{\sigma}{\lambda}\right).$$

Note that, in the Bayesian paradigm, it is straightforward to treat λ as a random variable by incorporating it into the hierarchical structure as a hyperprior, allowing it to be simultaneously estimated from the data.

As shown by [van Erp et al. \(2018\)](#), using a similar hierarchical construction we can obtain a Bayesian representation of the Ridge regression. In this case, each coefficient follows a normal distribution with a common variance governed by a global shrinkage parameter λ :

$$(2.9) \quad \begin{aligned} \beta_j \mid \lambda, \sigma^2 &\sim \mathcal{N}\left(0, \frac{\sigma^2}{\lambda}\right), \quad j = 1, \dots, p, \\ \lambda &\sim \text{half-Cauchy}^+(0, 1). \end{aligned}$$

[Hsiang \(1975\)](#) and [van Erp et al. \(2018\)](#) demonstrate that the posterior mean obtained from this hierarchical Ridge model leads to an ℓ_2 -type penalty, consistent with the form used in classical Ridge regression. In this interpretation, the half-Cauchy prior effectively induces a penalty of the form $|\beta_j|^p$ with $p = 2$.

The Elastic Net estimator also can be understood in a Bayesian framework as a maximum a posteriori (MAP) estimate where the prior on the coefficients β_j combines properties of both the Laplace and Gaussian distributions. The Elastic Net optimization problem is equivalent to MAP estimation under the following prior:

$$(2.10) \quad p(\beta_j) \propto \exp\left(-\lambda_1|\beta_j| - \lambda_2\beta_j^2\right).$$

This prior is a product of a Laplace component, which encourages sparsity, and a Gaussian component, which encourages shrinkage. Although this is not a standard probability distribution, it effectively captures the regularization behavior of both LASSO and ridge regression. Thus, the Elastic Net provides a Bayesian compromise between variable selection and coefficient stabilization.

2.4.2. HorseShoe prior

The Horseshoe prior, introduced by [Carvalho et al. \(2010\)](#), is one of the most widely used shrinkage priors in the Bayesian literature. A key property of this prior is its global–local adaptability. Unlike the shrinkage methods previously discussed, the Horseshoe prior induces sparsity through a hierarchical structure involving two levels of shrinkage: a global parameter that controls overall shrinkage across all coefficients, and local parameters that act individually on each regression coefficient.

As discussed by [Carvalho et al. \(2010\)](#), this structure allows the Horseshoe prior to capture many of the benefits of discrete mixture models (such as spike-and-slab) while avoiding their computational complexity.

The hierarchical form of the Horseshoe prior is given by:

$$(2.11) \quad \begin{aligned} \beta_j \mid \lambda_j, \tau &\sim \mathcal{N}(0, \lambda_j^2 \tau^2), \\ \lambda_j &\sim \mathcal{C}^+(0, 1), \\ \tau &\sim \mathcal{C}^+(0, 1), \end{aligned}$$

where $\mathcal{C}^+(0, 1)$ denotes the standard half-Cauchy distribution. Here, τ is the global shrinkage parameter, shared across all coefficients, while each λ_j serves as a local shrinkage parameter specific to coefficient β_j .

According to [Carvalho et al. \(2009\)](#), this configuration allows the model to strongly shrink coefficients associated with noise (i.e., values of β_j close to zero), while permitting large coefficients to escape shrinkage due to the heavy-tailed nature of the half-Cauchy distribution. In other words, the Horseshoe prior provides strong shrinkage near the origin and weak shrinkage in the tails, thereby offering an effective and robust mechanism for handling sparsity.

2.4.3. Spike-and-Slab priors

The stochastic search variable selection (SSVS), which uses Spike-and-Slab priors, as proposed by [George and McCulloch \(1993\)](#), introduces a Bayesian variable selection strategy through a hierarchical prior structure that combines a point-mass "spike" with a more dispersed "slab" distribution into a mixture. Rather than fixing the model size or selecting a subset directly, their approach introduces a latent binary indicator vector $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_p)$, where each $\gamma_j \in \{0, 1\}$ determines whether the j -th regressor is included in the model.

The regression model is given by:

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n),$$

with a prior on $\boldsymbol{\beta}$ conditional on $\boldsymbol{\gamma}$ defined as:

$$(2.12) \quad \boldsymbol{\beta} \mid \boldsymbol{\gamma}, \sigma^2 \sim \mathcal{N}(\mathbf{0}, \sigma^2 D_{\boldsymbol{\gamma}}),$$

where $D_{\boldsymbol{\gamma}} = \text{diag}(d_1, \dots, d_p)$, with:

$$d_j = \begin{cases} \tau_1^2 & \text{if } \gamma_j = 1 \quad (\text{slab}) \\ \tau_0^2 & \text{if } \gamma_j = 0 \quad (\text{spike}) \end{cases}$$

and typically, $\tau_0^2 \ll \tau_1^2$ (the spike variance smaller than the slab variance), enforcing strong shrinkage toward zero when $\gamma_j = 0$.

The indicator variables γ_j follow independent Bernoulli priors:

$$\gamma_j \sim \text{Bernoulli}(\pi), \quad \text{for } j = 1, \dots, p,$$

with π either fixed or assigned a prior distribution (e.g., uniform or beta), depending on modeling choice. With this approach when $\gamma_j = 0$, the prior variance of β_j is very small (τ_0^2), concentrating the prior mass near zero and when $\gamma_j = 1$, β_j is allowed to vary freely under the broader variance τ_1^2 .

This structure facilitates automatic variable selection via posterior inference on $\boldsymbol{\gamma}$, and allows efficient computation via Gibbs sampling, as the conditional distributions are standard. Importantly, [George and McCulloch \(1993\)](#) show that this formulation performs Bayesian model averaging across different subsets defined by $\boldsymbol{\gamma}$, without requiring explicit enumeration of all 2^p models.

The spike-and-slab prior thus combines model uncertainty and regularization in a coherent Bayesian framework, and forms the basis for many modern extensions in high-dimensional inference and sparse modeling.

[Bai et al. \(2021\)](#) demonstrate that the mixture-based structure of spike-and-slab priors yields shrinkage behavior that is less sensitive to the strength of the penalization. Unlike continuous shrinkage priors such as the Lasso or Ridge, where the penalization is uniformly applied to all coefficients, spike-and-slab priors apply shrinkage through a latent mixture mechanism.

In particular, when the spike variance becomes sufficiently small (or equivalently, when the slab variance increases), the posterior distribution selectively shrinks only the inactive coefficients. As a result, active coefficients that are not shrunk by the spike component tend to exhibit reduced bias compared to estimators based on global penalties, such as those induced by Lasso or Ridge regression.

As discussed by [George and Ročková \(2020\)](#), the concept of Bayesian penalty mixing arises from viewing classical regularization methods, such as ridge regression and the Lasso, through a Bayesian lens. In this framework, each penalty function corresponds to the log-prior of a particular distribution over the regression coefficients. For instance, the ridge penalty

$$(2.13) \quad \text{pen}_{\text{ridge}}(\beta) = \lambda \sum_{j=1}^p \beta_j^2$$

is equivalent to placing an independent Gaussian prior $\beta_j \sim \mathcal{N}(0, 1/\lambda)$ on each coefficient, while the Lasso penalty

$$(2.14) \quad \text{pen}_{\text{lasso}}(\beta) = \lambda \sum_{j=1}^p |\beta_j|$$

corresponds to independent Laplace priors $\beta_j \sim \text{Laplace}(0, 1/\lambda)$ As shown in earlier sections of this paper.

A key insight is that these classical penalties are separable sums over j , reflecting an implicit assumption that all β_j are a-priori independent and identically distributed ([George and Ročková, 2020](#)).

Bayesian penalty mixing induced by the spike and slab specification, generalizes this idea by constructing a penalty function from a probabilistic mixture of two or more underlying priors. Concretely, let $\pi_{\lambda_0}(\beta_j)$ and $\pi_{\lambda_1}(\beta_j)$ denote two candidate priors (e.g., a “spike” prior highly concentrated at zero, and a “slab” prior more diffuse). For a fixed mixing weight $\theta \in [0, 1]$, one defines the mixture prior

$$(2.15) \quad \pi_{\text{mix}, \theta}(\beta) = \prod_{j=1}^p [\theta \pi_{\lambda_1}(\beta_j) + (1 - \theta) \pi_{\lambda_0}(\beta_j)],$$

and the associated penalty function

$$\text{pen}_{\text{mix}, \theta}(\beta) = \log \left[\frac{\pi_{\text{mix}, \theta}(\beta)}{\pi_{\text{mix}, \theta}(0)} \right].$$

By mixing in the space of priors rather than directly combining penalty values, the resulting $\text{pen}_{\text{mix}, \theta}$ inherits adaptive shrinkage properties from both components.

When π_{λ_0} is chosen to be a “spike” distribution (very tightly concentrated near zero) and π_{λ_1} a “slab” distribution (more diffuse), the mixture prior $\pi_{\text{mix}, \theta}$ encodes prior belief that some coefficients are effectively zero while others may be large. In practice, one often sets π_{λ_0} and π_{λ_1} to be either two Gaussian priors (yielding the original “Gaussian spike-and-slab”) or two Laplace priors (yielding the Spike-and-Slab Lasso, or SSL). For example, in the Spike-and-Slab LASSO of [Bai et al. \(2021\)](#) takes

$$\pi_{\lambda_0}(\beta_j) = \frac{\lambda_0}{2} e^{-\lambda_0 |\beta_j|}, \quad \pi_{\lambda_1}(\beta_j) = \frac{\lambda_1}{2} e^{-\lambda_1 |\beta_j|}, \quad \lambda_0 \gg \lambda_1.$$

The score function for the mixed penalty $\text{pen}_{\text{mix},\theta}$ admits an enlightening interpretation. Denote

$$\text{pen}_{\lambda_0}(\beta) = \log\left[\frac{\pi_{\lambda_0}(\beta)}{\pi_{\lambda_0}(0)}\right], \quad \text{pen}_{\lambda_1}(\beta) = \log\left[\frac{\pi_{\lambda_1}(\beta)}{\pi_{\lambda_1}(0)}\right].$$

Then, for each coordinate j , the partial derivative of $\text{pen}_{\text{mix},\theta}$ with respect to β_j is

$$\frac{\partial \text{pen}_{\text{mix},\theta}(\beta)}{\partial \beta_j} = p_{\theta}^*(\beta_j) \frac{\partial \text{pen}_{\lambda_1}(\beta)}{\partial \beta_j} + (1 - p_{\theta}^*(\beta_j)) \frac{\partial \text{pen}_{\lambda_0}(\beta)}{\partial \beta_j},$$

where

$$(2.16) \quad p_{\theta}^*(\beta_j) = \frac{\theta \pi_{\lambda_1}(\beta_j)}{\theta \pi_{\lambda_1}(\beta_j) + (1 - \theta) \pi_{\lambda_0}(\beta_j)}$$

is the posterior probability (under the mixture prior) that β_j was drawn from the slab component. In words, the shrinkage applied to β_j is an adaptive convex combination of the two component shrinkages: when $|\beta_j|$ is small, $p_{\theta}^*(\beta_j)$ is close to zero and the spike penalty dominates, but when $|\beta_j|$ is large, $p_{\theta}^*(\beta_j)$ is near one and the slab penalty takes over (Ročková and George, 2018). Works such as George and McCulloch (1993), Bai et al. (2021) and George and Ročková (2020) shows that this mechanism preserves strong thresholding of negligible coefficients while protecting large coefficients from excessive shrinkage.

The simulation experiment presented below is designed to illustrate the qualitative behavior of different shrinkage mechanisms rather than to represent a fully high-dimensional setting. In particular, we use a relatively small number of predictors in order to visualize the coefficient trajectories and clearly highlight the shrinkage patterns induced by each method.

Our primary focus in this paper is on reviewing methodological developments for high-dimensional time-varying parameter models, while the numerical illustrations serve only as pedagogical examples of how these priors behave in practice.

To make a comparison of these methods, and its properties, we begin by generating a synthetic dataset with n observations and $p = 12$ covariates. Of these twelve predictors, eight are truly irrelevant (coefficient exactly zero), while four carry nonzero signal. Concretely, we set

$$\beta_{\text{true}} = (0, 0, 1.05, 0, 0, 0.55, 0, 0, 0, -0.75, -1.25, 0),$$

so that only $\beta_3 = 1.05$, $\beta_6 = 0.55$, $\beta_{10} = -0.75$ and $\beta_{11} = -1.25$ are nonzero. The response vector y is then formed as $\mathbf{y} = \mathbf{X} \beta_{\text{true}} + \varepsilon$, with $\varepsilon \sim \mathcal{N}(\mathbf{0}, I)$. All covariates \mathbf{X} are drawn i.i.d. from $\mathcal{N}(0, 1)$, and we standardize \mathbf{X} to have mean zero and unit variance as well as center y . This setup allows us to compare how different shrinkage and spike-and-slab methods recover the four true signals while driving the remaining eight coefficients to zero. We detail the implementation details in the Appendix A.

In Figure 1, we illustrate how each method shrinks coefficients on a dataset of size $n = 50$. Because the sample is small, the spike component forces most coefficients very close to zero until the penalty weakens sufficiently for true signals to emerge. Notice how Lasso and Elastic Net gradually drive some coefficients exactly to zero, while Ridge merely shrinks them continuously. The EMVS and SSLASSO panels demonstrate that, as v_0 or λ_0 increase, the posterior mean transitions from near-zero (spike) to the slab level only when there is enough evidence in the data.

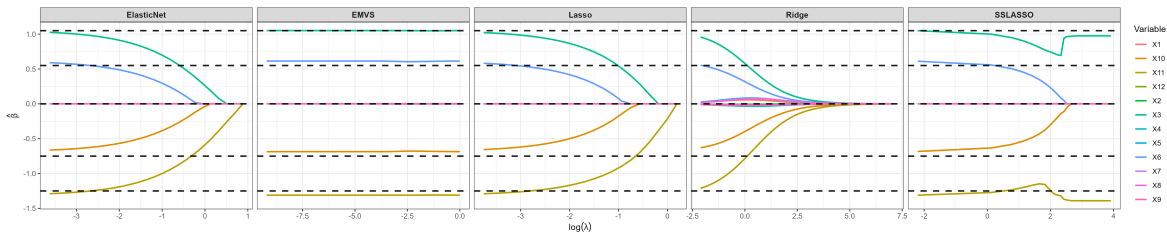


Figure 1: Coefficient trajectories ($\hat{\beta}_j$) versus $\log(\text{penalty})$ on a simulated dataset with $n = 50$. True coefficient values are marked by horizontal dashed lines.

In Figure 2, we repeat the same experiment with $n = 500$. With a larger sample, the mixture-type penalties clearly distinguish true signals from noise: the spike distribution no longer pulls strong coefficients to zero once there is substantial evidence, whereas the slab component allows those coefficients to remain near their true values. Thus, as data size increases, the posterior mean under EMVS and SSLASSO escapes spike-induced bias more rapidly, illustrating desirable asymptotic properties of spike-and-slab-mixture priors.

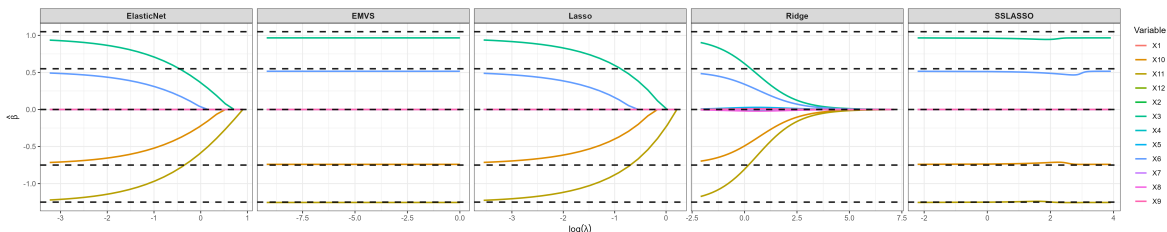


Figure 2: Coefficient trajectories ($\hat{\beta}_j$) versus $\log(\text{penalty})$ on a simulated dataset with $n = 500$. Horizontal dashed lines again mark the true coefficient values.

In addition to visual inspection of coefficient paths, several studies in the variable selection literature evaluate the performance of shrinkage methods using explicit selection criteria. For instance, [Yüzbaşı et al. \(2021\)](#) propose metrics such as the median model error, the average number of true zero coefficients correctly estimated as zero, and the average number of nonzero coefficients incorrectly estimated as zero. These criteria provide a more formal way of evaluating the variable selection ability of shrinkage methods and are frequently used in simulation studies involving high-dimensional regression. Although our numerical examples focus on illustrating shrinkage behavior rather than conducting a full performance comparison, these metrics represent useful tools for assessing variable selection accuracy in future empirical work.

3. DYNAMIC SHRINKAGE

Time-varying parameter (TVP) models are widely used in econometrics and time series analysis due to their ability to adapt to structural changes, regime shifts, and unexpected shocks. These models are particularly valuable for forecasting and inference in dynamic environments. However, a key limitation of TVP models is their tendency to overfit, especially when applied to high-dimensional datasets where the number of predictors may exceed the number of observations.

To address this challenge, a range of sparse modeling techniques has been developed. One major concern in high-dimensional settings is computational efficiency. While estimating TVP models via Gibbs sampling or other Markov Chain Monte Carlo (MCMC) methods is relatively straightforward for low-dimensional problems, their scalability is severely limited as the number of predictors grows. As noted by [Hauzenberger et al. \(2024\)](#), MCMC methods become computationally burdensome in high-dimensional settings, restricting their practical applicability.

In response to these challenges, a growing body of literature has focused on developing efficient estimation techniques for dynamic shrinkage. For example, [Ročková and McAlinn \(2021\)](#) propose an Expectation-Maximization (EM) algorithm for dynamic spike-and-slab priors, offering a scalable alternative to full Bayesian sampling. [Koop and Korobilis \(2023\)](#) and [Bernardi et al. \(2023\)](#) introduce a Variational Bayes (VB) approach that provides fast and approximate inference for dynamic shrinkage models.

Although significantly more challenging than in low-dimensional problems, it is also possible to employ the Markov Chain Monte Carlo (MCMC) algorithm for the estimation of dynamic sparsity models, albeit with certain caveats. [Hauzenberger et al. \(2024\)](#) address this by proposing dynamic horseshoe priors. This approach facilitates the temporal evolution of only a subset of coefficients, with changes occurring exclusively at specific, selected time points. In this section, we review recent contributions that propose innovative approaches for modeling dynamic sparsity, with a focus on both computational efficiency and inferential performance.

3.1. Dynamic spike-and-slab

3.1.1. Expectation maximization approach

The framework proposed by [Ročková and McAlinn \(2021\)](#), based on an Expectation-Maximization (EM) algorithm, can be described as follows. Consider a regression model that adopts a dynamic spike-and-slab structure, such that:

$$(3.1) \quad y_t = \beta_{t,j} x_{t-h} + \varepsilon_t,$$

where

$$\begin{aligned} \varepsilon_t &\sim N(0, \sigma), \\ \beta_{t,j} &\sim N(\gamma_{t,j} \mu_{t,j}, \gamma_t \lambda_1 + (1 - \gamma_t) \lambda_0), \\ \mu_{t,j} &= \phi_0 + \phi \mu_{t-1,j} + \nu, \quad \nu \sim N(0, \sigma_\mu), \\ \gamma_{t,j} &= \begin{cases} 0 & \text{if spike,} \\ 1 & \text{if slab.} \end{cases} \end{aligned}$$

Here, λ_1 and λ_0 represent the variances of each component of the mixture of the Dynamic Spike-and-Slab, the Spike and The Slab, respectively. Throughout the paper we use $\gamma_{j,t} = 1$ to denote inclusion (slab) and $\gamma_{j,t} = 0$ to denote exclusion (spike).

The choice of ϕ is made such that $|\phi| < 1$, ensuring stationarity of the parameters. This allows for a structure such that μ_t becomes a stationary process with mean ϕ_0 and variance $\frac{\lambda_1}{1-\phi^2}$.

The inference procedure employed in this study is based on *Maximum a Posteriori* (MAP) estimation, a Bayesian approach to parameter inference. MAP estimation identifies the parameter θ that maximizes the posterior probability $P(\theta | X)$, given the observed data X . Formally, it is defined as:

$$\hat{\theta}_{\text{MAP}} = \arg \max_{\theta} P(\theta | X).$$

Using Bayes' theorem:

$$P(\theta | X) = \frac{P(X | \theta)P(\theta)}{P(X)}.$$

Since $P(X)$ is a constant with respect to θ , MAP estimation simplifies to:

$$\hat{\theta}_{\text{MAP}} = \arg \max_{\theta} P(X | \theta)P(\theta)$$

where $P(X | \theta)$ is the likelihood of the data given the parameter and $P(\theta)$ is the prior distribution of the parameter.

MAP estimation incorporates prior knowledge, which helps in regularization and prevents overfitting. Unlike Maximum Likelihood Estimation (MLE), which only considers the likelihood function, MAP includes prior information, making it more robust, particularly in small-sample settings. The prior acts as a form of regularization, ensuring that the estimates remain stable. This makes MAP especially useful when data is limited or when MLE produces ill-posed solutions.

To derive the MAP (Maximum A Posteriori) estimator for $\beta_{1:T,j}$, [Ročková and McAlinn \(2021\)](#) maximize the posterior distribution:

$$(3.2) \quad \hat{\beta}_{1:T,j} = \arg \max_{\beta_{1:T,j}} \pi(\beta_{1:T,j} | Y_{1:T}),$$

where $\pi(\beta_{1:T,j} | Y_{1:T})$ is proportional to the product of the likelihood and the prior:

$$(3.3) \quad \pi(\beta_{1:T,j} | Y_{1:T}) \propto L(Y_{1:T} | \beta_{1:T,j}) \cdot \pi(\beta_{1:T,j}).$$

The prior for $\beta_{t,j}$ is defined as a mixture of the spike and slab distributions:

$$(3.4) \quad \pi(\beta_{t,j} | \gamma_{t,j}, \beta_{t-1,j}) = (1 - \gamma_{t,j})\psi_0(\beta_{t,j} | \lambda_0) + \gamma_{t,j}\psi_1(\beta_{t,j} | \mu_t, \lambda_1),$$

where:

- $\psi_0(\beta_t | \lambda_0)$ is the spike distribution with variance λ_0 and mean zero,
- $\psi_1(\beta_t | \mu_t, \lambda_1)$ is the slab distribution with mean μ_t and variance λ_1 ,
- $\gamma_t \in \{0, 1\}$ is the binary indicator for spike/slab membership.

[Ročková and McAlinn \(2021\)](#) assume Gaussian distributions for the slab and spike components, denoted by ψ_1 and ψ_0 , respectively. They define the inclusion parameter γ_{t_j} according to the following structure:

$$(3.5) \quad P(\gamma_t = 1 | \beta_{t-1}) = \theta_t,$$

where $\gamma_t = 1$ denotes a slab (included), so θ_t represents the prior probability of inclusion.

The specification given by [Ročková and McAlinn \(2021\)](#) for θ_t involves introducing an interpretable parameter called the marginal importance weight Θ , where $0 \leq \Theta \leq 1$ controls the overall balance between the spike and slab distributions.

Given $(\Theta, \lambda_0, \lambda_1, \phi_0, \phi_1)$, the conditional inclusion probability θ_t (or a transition function $\theta(\beta_{t-1})$) is defined as:

$$(3.6) \quad \theta_t \equiv \theta(\beta_{t-1}) = \frac{\Theta \psi_1^{ST}(\beta_{t-1} | \lambda_1, \phi_0, \phi_1)}{\Theta \psi_1^{ST}(\beta_{t-1} | \lambda_1, \phi_0, \phi_1) + (1 - \Theta) \psi_0(\beta_{t-1} | \lambda_0)}.$$

It is important to note that the values of Θ correspond to two limiting cases of the model. When $\Theta = 1$, the model reduces to a standard dynamic linear model (DLM) without any shrinkage. In contrast, when $\Theta = 0$, the model relies exclusively on the spike distribution, resulting in all parameters being shrunk toward zero.

3.1.2. Variational Bayes approach

[Koop and Korobilis \(2023\)](#) propose a novel Variational Bayes algorithm for dynamic variable selection in time-varying parameter regression models with a large number of predictors. The method identifies, at each point in time, which covariates are relevant for forecasting the dependent variable. Applied to the problem of forecasting inflation using over 400 macroeconomic, financial, and global indicators, many of which may be irrelevant or short-lived, the approach delivers sparse, parsimonious solutions and demonstrates strong predictive performance in a high-dimensional setting.

A different approach is adopted compared to [Ročková and McAlinn \(2021\)](#). Instead of using the Expectation-Maximization (EM) algorithm, the variational Bayes method seeks to approximate the true posterior distribution by minimizing the Kullback–Leibler (KL) divergence between the variational density and the posterior ([Koop and Korobilis, 2023](#)). The KL divergence is defined as:

$$(3.7) \quad KL(q || p) = \int q(s, \theta | y) \log \left\{ \frac{q(s, \theta | y)}{p(s, \theta | y)} \right\} ds d\theta.$$

The goal is to find the optimal variational density $q^*(s, \theta | y)$ that minimizes the KL divergence:

$$(3.8) \quad q^*(s, \theta | y) = \arg \min_{q(s, \theta | y) \in \mathcal{F}} KL(q || p),$$

with \mathcal{F} denoting the family of admissible variational densities. This optimization yields a tractable approximation to the posterior, balancing computational efficiency and accuracy in complex dynamic models.

As shown by [Nakajima et al. \(2019\)](#) and [Koop and Korobilis \(2023\)](#), the mean-field factorization of the variational distribution, given by $q(s, \theta | y) = q(\theta | y)q(s | y)$, enables

an efficient iterative optimization scheme. By conditioning on one set of variables while optimizing over the other, the algorithm alternates between the latent parameter blocks, simplifying the overall computation.

Since the Kullback–Leibler divergence $KL(q || p)$ is always non-negative and equals zero only when the variational distribution exactly matches the true posterior, i.e., $q(s, \theta | y) = p(s, \theta | y)$ (Koop and Korobilis, 2023), the algorithm is considered to have converged once changes in the objective function (Evidence Lower Bound, ELBO) falls below a pre-specified threshold.

The model proposed by Koop and Korobilis (2023) adopts a dynamic extension of the stochastic search variable selection (SSVS) framework originally introduced by George and McCulloch (1993). In this specification, the regression coefficients follow a mixture of two normal distributions, where a binary indicator changing in time governs whether each coefficient is drawn from a narrow “spike” distribution or a wider time varying “slab” distribution. The hierarchical prior structure is defined as follows:

$$\begin{aligned}\beta_{j,t} | \gamma_{j,t}, \tau_{j,t}^2 &\sim (1 - \gamma_{j,t})\mathcal{N}\left(0, c \times \tau_{j,t}^2\right) + \gamma_{j,t}\mathcal{N}\left(0, \tau_{j,t}^2\right), \\ \gamma_{j,t} | \pi_{0,t} &\sim \text{Bernoulli}(\pi_{0,t}), \\ \frac{1}{\tau_{j,t}^2} &\sim \text{Gamma}(g_0, h_0), \\ \pi_{0,t} &\sim \text{Beta}(1, 1),\end{aligned}$$

where $\gamma_{j,t} \in \{0, 1\}$ is the dynamic inclusion indicator, $\tau_{j,t}^2$ is a local variance parameter, c is a small constant (typically $c \ll 1$) controlling the spike’s tightness, and $\pi_{0,t}$ governs the inclusion probability across time. This formulation allows the degree of shrinkage and variable inclusion to vary dynamically over time, promoting sparsity in a flexible and time-adaptive manner.

An important feature of the dynamic variable selection prior used by Koop and Korobilis (2023), as in the Ročková and McAlinn (2021) model, is that it produces time-varying posterior inclusion probabilities (PIPs). Under their specification, the prior mean inclusion probability at time t is given by $\mathbb{E}(p(\pi_{0,t})) = \frac{1}{2}$, while the posterior mean of the inclusion indicator $\gamma_{j,t}$, denoted $\tilde{\pi}_{j,t} = \mathbb{E}(\gamma_{j,t} | y_{1:T})$, reflects the time-specific PIP for predictor j . This formulation generalizes the fixed PIP concept from George and McCulloch (1993) to a dynamic setting by allowing the hyperparameters π_t , τ_t^2 , and γ_t to evolve over time.

The full prior distribution of the model is given by:

$$(3.9) \quad \begin{aligned}p(\beta_{1:T}, \sigma_{1:T}^2, \mathbf{w}_{1:T}, \gamma_{1:T}, \tau_{1:T}^2, \pi_{0,1:T}) &= \prod_{t=1}^T p(\beta_t | \beta_{t-1}, w_t) p(w_t) p(\sigma_t^2) \\ &\times p(\beta_t | \gamma_t, \tau_t^2) p(\tau_t^2) p(\gamma_t | \pi_{0,t}) p(\pi_{0,t})\end{aligned}$$

where the first term is a Gaussian TVP prior for β_t , w_t governs state equation variances with an inverse gamma prior, σ_t^2 is a stochastic volatility parameter, the remaining terms follow the dynamic variable selection (DVS) structure introduced earlier.

To simplify inference, the mixture prior $p(\beta_t \mid \gamma_t, \tau_t^2)$ is re-expressed as a single-component normal distribution:

$$(3.10) \quad p(\beta_t \mid V_t) \sim \prod_{j=1}^p \mathcal{N}(0, v_{j,t}),$$

where each variance component is defined as:

$$v_{j,t} = (1 - \gamma_{j,t})^2 c \times \tau_{j,t}^2 + \gamma_{j,t}^2 \tau_{j,t}^2,$$

and $V_t = \text{diag}(v_{1,t}, \dots, v_{p,t})$ is a diagonal matrix, and c controls the spike component.

This reparameterization allows the posterior for $\beta_{1:T}$ to be efficiently computed within the variational inference framework, exploiting the conditional independence and hierarchical structure of the DVS prior.

By applying this formulation, [Koop and Korobilis \(2023\)](#) derive the variational Bayes posterior distribution for the time-varying coefficients $\beta_{1:T}$, conditional on the observed data $y_{1:T}$. The variational Bayes posterior approximation takes the following form:

$$(3.11) \quad q(\beta_{1:T} \mid y_{1:T}) \propto \exp \left[\mathbb{E}_{q(\sigma_{1:T}^2 \mid y_{1:T})} \left(\sum_{t=1}^T \log p(y_t \mid \beta_t, \sigma_t^2) \right) \right. \\ \left. + \mathbb{E}_{q(w_{1:T} \mid y_{1:T})} \left(\sum_{t=1}^T \log p(\beta_t \mid \beta_{t-1}, w_t) \right) \right. \\ \left. + \mathbb{E}_{q(V_{1:T} \mid y_{1:T})} \left(\sum_{t=1}^T \log p(\beta_t \mid V_t) \right) \right].$$

Following the application of variational Bayes (VB) updates to the transformed state-space model, [Koop and Korobilis \(2023\)](#) derive smoothed posterior estimates for the state vector $\mathbf{m}_{1:T}$. These estimates are then used to update the dynamic variable selection (DVS) prior parameters $\tau_{j,t}^2$, $\gamma_{j,t}$, $v_{j,t}$, and $\pi_{0,t}$, based on conditional expectations with respect to the variational posterior $q(\beta_{1:T} \mid y_{1:T})$. The update steps are as follows:

$$(33) \quad \tau_{j,t}^2 = \mathbb{E} [q(\tau_{j,t}^2 \mid y)] = [h_0 + (m_{j,t|T}^2 + P_{j,j,t|T}) / 2] / [g_0 + 1/2],$$

$$(34) \quad \hat{\gamma}_{j,t} = \mathbb{E} [q(\gamma_{j,t} \mid y)] = \frac{\mathcal{N}(m_{j,t|T} \mid 0, \tau_{j,t}^2) \hat{\pi}_{0,t}}{\mathcal{N}(m_{j,t|T} \mid 0, \tau_{j,t}^2) \hat{\pi}_{0,t} + \mathcal{N}(m_{j,t|T} \mid 0, c \times \tau_{j,t}^2) (1 - \hat{\pi}_{0,t})},$$

$$(35) \quad \hat{v}_{j,t} = \mathbb{E} [q(v_{j,t} \mid y)] = (1 - \hat{\gamma}_{j,t})^2 c \hat{\tau}_{j,t}^2 + \hat{\gamma}_{j,t}^2 \hat{\tau}_{j,t}^2,$$

$$(36) \quad \hat{\pi}_{0,t} = \mathbb{E} [q(\pi_{0,t} \mid y)] = \left(1 + \sum_{j=1}^p \hat{\gamma}_{j,t} \right) / (2 + p).$$

These updates are computed for each $t = 1, \dots, T$ and $j = 1, \dots, p$, where all expectations $\mathbb{E}[\cdot]$ are taken with respect to the variational posterior distributions of the corresponding parameters.

3.1.3. Advances in variational Bayes for variable selection

Bernardi et al. (2023) propose a dynamic Bernoulli–Gaussian (BG) regression model for time-varying parameter estimation with dynamic variable selection in high-dimensional settings. The model adopts a variational Bayes approach and extends the foundational work of Koop and Korobilis (2023) by introducing several methodological innovations to address limitations such as temporal independence in inclusion decisions and noisy posterior inclusion probabilities.

The central idea is to model both the regression coefficients and their inclusion indicators as time-varying processes. The observation equation is specified as:

$$(3.12) \quad y_t = \sum_{j=1}^p \beta_{j,t} x_{j,t-1} + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, e^{h_t}),$$

where h_t is a latent log-volatility process and $x_{j,t-1}$ denotes the j -th predictor.

Each time-varying coefficient $\beta_{j,t}$ is modeled as the product of a latent state $b_{j,t}$ and a binary inclusion indicator $\gamma_{j,t}$:

$$(3.13) \quad \beta_{j,t} = b_{j,t} \cdot \gamma_{j,t}, \quad b_{j,t} = b_{j,t-1} + v_{j,t}, \quad v_{j,t} \sim \mathcal{N}(0, \eta_j^2).$$

The inclusion indicators $\gamma_{j,t} \in \{0, 1\}$ are Bernoulli-distributed with time-varying inclusion probabilities governed by a latent process:

$$(3.14) \quad \gamma_{j,t} \mid \omega_{j,t} \sim \text{Bernoulli}(\text{expit}(\omega_{j,t})), \quad \omega_j \sim \mathcal{N}(0, \xi_j^2 Q^{-1}),$$

where $\text{expit}(\omega_{j,t}) = \frac{1}{1 + \exp(-\omega_{j,t})}$, and Q is a tridiagonal precision matrix that encodes temporal dependence via a Gaussian Markov Random Field (GMRF) prior.

A key contribution of the model lies in the structured factorization of the variational posterior distribution. The authors allow predictors to be grouped into K latent clusters based on shared characteristics, enabling flexible modeling of inter-variable dependence. The variational distribution over the latent coefficients is factorized as:

$$q(\mathbf{b}) = \prod_{k=1}^K q(\mathbf{b}_k),$$

where \mathbf{b}_k is the vector of latent states for group k . This framework spans from the fully correlated case ($K = 1$) to the fully independent case ($K = p$), with intermediate block-correlation structures.

The full variational approximation is expressed as:

$$q(\vartheta) = q(\mathbf{h})q(\nu^2) \prod_{k=1}^K q(\mathbf{b}_k) \prod_{j=1}^p q(\omega_j)q(\eta_j^2)q(\xi_j^2) \prod_{t=1}^n q(\gamma_{j,t})q(z_{j,t}),$$

where \mathbf{h} is the log-volatility process, ω_j the latent logits for inclusion, η_j^2 and ξ_j^2 the local scale parameters, and $z_{j,t}$ auxiliary variables introduced via data augmentation.

This grouped factorization improves the balance between flexibility and computational tractability: it permits correlation among predictors within a group, while maintaining the scalability advantages of variational inference. Moreover, it allows one to incorporate domain knowledge to define meaningful predictor groupings.

To mitigate the noisiness of posterior inclusion probabilities $\mu_q(\gamma_{j,t})$, the authors introduce a smoothing step using B-spline basis functions. Specifically, the smoothed posterior inclusion probability is modeled as:

$$(3.15) \quad \tilde{\mu}_q(\gamma_{j,t}) = \text{expit}(w_t^\top f_j),$$

where w_t is the B-spline basis vector and f_j the coefficient vector for variable j . The smoothed version is obtained by minimizing the Kullback–Leibler divergence between the smoothed and original variational distributions. [Bernardi et al. \(2023\)](#) show that this procedure reduces noise in parameter estimates and mitigates erratic behavior in the inclusion trajectories.

Finally, the authors distinguish three versions of the BG model based on the structure of the factorization:

- **BG**: independent inclusion processes for each predictor ($K = p$);
- **BG joint**: a single inclusion process shared by all predictors ($K = 1$);
- **BG group**: block-wise inclusion processes shared within $K < p$ groups.

These variants differ only in the way the latent states \mathbf{b}_j are grouped in the variational factorization, allowing the modeler to control the degree of sparsity correlation across variables.

3.1.4. Models comparison

In [Bernardi et al. \(2023\)](#) empirical application, alternative models are compared by evaluating the out-of-sample forecast¹ performance for U.S. inflation. The comparison is based on the Root Mean Squared Forecast Error (RMSFE) across multiple forecast horizons and inflation measures, including CPI, core CPI, PCE, and the GDP deflator. The authors estimate the following three variants of their model and compares with other dynamic specifications:

- **TVI** ([Stock and Watson, 2007](#)): Local-level model with stochastic volatility.

¹The authors use the following specification for forecasting each variable:

$$y_{t+h} = \frac{400}{h} \ln\left(\frac{P_t}{P_{t-1}}\right).$$

It is worth noting that, as h increases, the variance of the target variable y_{t+h} decreases. This reduction in variance can create the appearance of better forecast accuracy at longer horizons, but it is actually a feature of the target’s construction rather than a genuine improvement in model performance. Nevertheless, comparing models on this common target remains useful for assessing relative forecasting ability.

- **AR(2)**: autoregressive model with two lags and no time-varying parameters.
- **TVAR(2)**: AR(2) model with time-varying coefficients and stochastic volatility (Koop and Korobilis, 2023).
- **BG (independent)**: Each predictor has its own independent latent inclusion probability process.
- **BG joint**: All predictors share a common inclusion probability process, implying complete pooling.
- **BG group**: The predictors are grouped and each group shares a common inclusion process, allowing for correlation within the groups but independence between groups.
- **DVS**: Dynamic Variable Selection with dynamic spike-and-slab prior Koop and Korobilis (2023), estimated via variational Bayes.
- **DSS**: Dynamic Spike-and-Slab model (Ročková and McAlinn, 2021), using a dynamic prior with MCMC-based inference.
- **DEMVS²**: Dynamic Expectation-Maximization Variable Selection Ročková and McAlinn (2021), spike-and-slab prior estimated via EM-algorithm.

Following the same empirical setup, we include forecasting results for the dynamic variant of the Expectation Maximization algorithm proposed by Ročková and McAlinn (2021), which serves as another contender for assessing the forecasting performance from structured dynamic sparsity.

At the one-step-ahead horizon, the simple time-varying intercept model (TVI) yields an average RMSFE of 1.59, while an AR(2) specification and a time-varying AR(2) (TVAR(2)) achieve averages of 1.61 and 1.62, respectively. Among Bayesian group-shrinkage priors, the Bayesian Grouped Spike-and-Slab (BG group) attains the lowest average error (1.54), closely followed by the basic Bayesian Spike-and-Slab (BGS) at 1.56. Our Dynamic EMVS (DEMVS) procedure posts an average RMSFE of 1.74; although this is marginally higher than BG group at $h = 1$, DEMVS still produces the smallest one-step-ahead error for GDPCTPI (1.50) and one of the two lowest errors for PCECTPI (1.73). These results suggest that, even on the shortest horizon, BG group is the best overall, while DEMVS’s mixture-type prior successfully isolates the four true predictive signals (e.g., GDPCTPI and CPIAUSCL) and remains highly competitive.

Moving to the two-step-ahead horizon, one observes a similar pattern. The AR(2) and TVAR(2) models report average RMSFEs of 1.46 and 1.26, respectively, whereas BGS and BG group reduce errors to around 1.00–1.05. DEMVS achieves an average RMSFE of 1.11, tying for the lowest error on CPIAUSCL (1.45) and CPILFESL (0.95), and yielding the smallest RMSFE on GDPCTPI (0.93). In this case, BG group slightly outperforms DEMVS overall at $h = 2$, but DEMVS remains the top model for GDPCTPI.

For $h = 4$, DEMVS attains an average RMSFE of 0.64, outperforming TVI (0.70), AR(2) (1.32), and even the best Bayesian group method (BGS at 0.66). In particular,

²Parameters used in the DEMVS model were $\phi_0 = 0$, $\phi_1 = 0.98$, $\lambda_0 = 10 \cdot (1 - 0.98^2)$, $\lambda_1 = 0.9$, and $\Theta = 0.96$.

DEMVS’s RMSFEs for GDPCTPI (0.59) and CPILFESL (0.60) are less than half of those from static or simpler dynamic priors. At $h = 8$, DEMVS’s average RMSFE falls to 0.35, compared to 0.37 for TVAR(2) and 0.36 for BGH. The error on GDPCTPI (0.32) is especially notable—roughly half of that achieved by Lasso-type dynamic shrinkage, the DSS(5) or DSS(9). Thus, DEMVS emerges as the best-performing model at longer horizons ($h \geq 4$), illustrating that its spike-and-slab mixture prior allows the posterior mean to escape spike-induced bias and concentrate on true signals as more data accumulate.

Overall, the BG group model delivers the best forecasting accuracy at short and medium horizons ($h = 1, 2$), suggesting that modeling structured sparsity through groupwise dependence effectively balances flexibility and parsimony. At longer horizons ($h = 4, 8$), however, DEMVS takes the lead. Consequently, DEMVS and BG models perform competitively across all horizons and stand out among the methods compared.

3.2. Scalable markov chain monte carlo

The computational demands of Markov Chain Monte Carlo (MCMC) methods limit their applicability to cases where the number of predictors is not too large (Hauzenberger et al., 2024). Nonetheless, MCMC (with some adaptations) remains viable up to a certain scale. In light of these challenges, Hauzenberger et al. (2024) propose a new dynamic shrinkage prior that captures the empirical regularity that time-varying parameters (TVPs) are typically sparse, in the case where time variation occurs only sporadically and only for a subset of coefficients.

The dynamic horseshoe approach of Hauzenberger et al. (2024) is motivated by the empirical observation that time variation in regression coefficients is often sparse: only a small subset of predictors experiences meaningful changes over time, while most coefficients remain approximately constant. Instead of allowing each coefficient to evolve freely at every time step, the model introduces shrinkage through a hierarchical horseshoe prior that strongly penalizes unnecessary time variation. By allowing the global shrinkage parameter to vary over time while keeping local shrinkage parameters static, the model captures occasional bursts of parameter instability while maintaining computational tractability in high-dimensional settings.

The authors consider a static representation of a TVP regression model involving a T -dimensional dependent variable, \mathbf{y} , and a $T \times K$ matrix of predictors, \mathbf{X} :

$$(3.16) \quad \mathbf{y} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{W}\boldsymbol{\beta} + \mathbf{L}\boldsymbol{\epsilon}$$

$$\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}_T), \quad \boldsymbol{\beta} = (\boldsymbol{\beta}'_1, \dots, \boldsymbol{\beta}'_T)'$$

Here, $\boldsymbol{\alpha}$ is a K -dimensional vector of time-invariant coefficients, $\boldsymbol{\beta}_t$ is a $K \times 1$ vector of time-varying coefficients, and $\mathbf{L} = \text{diag}(\sigma_1, \dots, \sigma_T)$, where σ_t denotes time-varying error volatilities. The TVP structure arises from the term $\mathbf{W}\boldsymbol{\beta}$, where \mathbf{W} is a $T \times k$ matrix.

To handle the high-dimensional nature of this model, the authors adopt the scalable MCMC strategy proposed by Johndrow et al. (2018). An exact MCMC algorithm is used for

α , while an approximation is employed for the typically sparse and high-dimensional β_t . The core idea is to approximate the high-dimensional matrix Γ by dropping irrelevant columns of $\widetilde{\mathbf{W}}$, thereby improving computational efficiency.

The algorithm proceeds as follows:

1. Draw a k -dimensional vector $v \sim \mathcal{N}(0_k, \mathbf{D}_0)$,
2. Sample a T -dimensional vector $q \sim \mathcal{N}(0_T, \mathbf{I}_T)$,
3. Define $w = \widetilde{\mathbf{W}}v + q$,
4. Solve $(\tilde{\mathbf{y}} - w) = (\mathbf{I}_T + \hat{\Gamma})u$ for u , with $\hat{\Gamma} = \widetilde{\mathbf{W}}_S \mathbf{D}_{0,S} \widetilde{\mathbf{W}}_S'$,
5. Set $\beta = \mathbf{D}_{0,S} \widetilde{\mathbf{W}}_S' u + v$.

In this setting, $\widetilde{\mathbf{W}}_S$ denotes the $T \times s$ submatrix of $\widetilde{\mathbf{W}}$ composed of columns indexed by the set S , and $\mathbf{D}_{0,S}$ is the corresponding submatrix of \mathbf{D}_0 . The set $S = \{j : \delta_j = 1\}$ is defined by a binary selection vector δ , with $\delta_j = 1$ with probability p_j and $\delta_j = 0$ with probability $1 - p_j$. [Hauzenberger et al. \(2024\)](#) employ the Signal Adaptive Variable Selection (SAVS) method to approximate δ , which is then used in the MCMC scheme.

The authors specify a horseshoe prior for β_t as follows:

$$(3.17) \quad p(\beta_t) = \prod_{j=1}^K \mathcal{N}(\beta_{jt} \mid 0, \tau \lambda_t \phi_{jt}^2), \quad \phi_{jt} \sim \mathcal{C}^+(0, 1),$$

where $\beta_t = (\beta_{1t}, \dots, \beta_{Kt})'$ is the vector of time-varying coefficients at time t , τ is a global shrinkage parameter, λ_t is a time-specific global shrinkage factor, and ϕ_{jt} are local shrinkage parameters with half-Cauchy priors.

Importantly, the [Hauzenberger et al. \(2024\)](#) introduce temporal dynamics only in the common global shrinkage factor λ_t , which simplifies the model and reduces computational costs. By omitting the time evolution of each individual local shrinkage parameter $\phi_{j,t}$, the model leverages cross-sectional information to perform shrinkage more efficiently while retaining flexibility.

To estimate λ_t under an AR(1) process, the authors propose an approximation that renders the hierarchical model linear and conditionally Gaussian. They define the transformed coefficients as:

$$(3.18) \quad \hat{\beta}_t = \frac{\beta_t}{\phi_{jt} \sqrt{\tau}},$$

and construct the scalar quantity $b_t = \hat{\beta}_t^\top \hat{\beta}_t = \lambda_t \nu_t$, where $\nu_t = \mathbf{v}_t^\top \mathbf{v}_t \sim \chi_K^2$. Since sampling from the χ_K^2 distribution is computationally costly in high-dimensional contexts, the authors use the Gaussian approximation $\nu_t \approx \sqrt{2K} q_t + K$, where $q_t \sim \mathcal{N}(0, 1)$. This leads to a Gaussian approximation for $\log b_t = \log \lambda_t + \log \nu_t$.

For large values of K ($K > 50$), the approximation is highly accurate and enables the use of standard Gaussian linear state-space methods to sample $\log \lambda_t$. Additionally,

the algorithm constructs the variance matrix as $D_0 = \text{diag}(\Omega_1, \dots, \Omega_T)$ and samples the remaining parameters using standard Bayesian updating techniques.³

4. CONCLUSION

This survey has reviewed recent advances in dynamic shrinkage models for time-varying parameter estimation and variable selection, with a focus on macroeconomic and financial time series. Building on the foundational work in static regularization—such as Ridge (Hoerl and Kennard, 2000), Lasso (Tibshirani, 1996), and Bayesian priors like the horseshoe (Carvalho et al., 2010) and spike-and-slab (George and McCulloch, 1993), we examined how these ideas extend to dynamic frameworks that accommodate structural change and time-varying sparsity. These innovations are essential in modern forecasting settings where economic relationships evolve, and the set of relevant predictors is not static.

We highlighted several methodological approaches, including Expectation Maximization (Ročková and McAlinn, 2021) and Variational Bayes algorithms for dynamic spike-and-slab models (Koop and Korobilis, 2023; Bernardi et al., 2023), as well as scalable MCMC strategies for dynamic horseshoe priors (Hauzenberger et al., 2024). Each method addresses the twin challenges of flexibility and scalability in distinct ways. The comparison of empirical results in Bernardi et al. (2023) is augmented by our application on the dynamic version of the EMVS algorithm of Ročková and McAlinn (2021), particularly in the context of U.S. inflation forecasting, illustrates that structured priors—such as those used in the BG group model and the dynamic horseshoe—yield significant performance gains by effectively balancing sparsity and adaptiveness.

Despite these developments, several open questions remain. Future research should explore unified frameworks that combine global-local shrinkage, time-varying inclusion probabilities, and model-based grouping strategies in a computationally efficient manner. In addition, more work is needed to evaluate the robustness of dynamic shrinkage methods under model misspecification. Overall, the growing body of literature of dynamic shrinkage techniques offers promising paths forward for high-dimensional inference in evolving economic, financial, and time series environments.

ACKNOWLEDGMENTS

This research was funded by Capes (Finance Code 01) , CNPq (310646/2021-9) and FAPESP (2023/02538-0).

³As detailed in the appendix of Hauzenberger et al. (2024).

REFERENCES

- Bai, R., Ročková, V., and George, E. I. (2021). *Spike-and-Slab Meets LASSO: A Review of the Spike-and-Slab LASSO*, page 81–108. Chapman and Hall/CRC.
- Bernardi, M., Bianchi, D., and Bianco, N. (2023). Dynamic variable selection in high-dimensional predictive regressions. Available at SSRN: <https://ssrn.com/abstract=4418264>.
- Carvalho, C. M., Polson, N. G., and Scott, J. G. (2009). Handling sparsity via the horseshoe. In van Dyk, D. and Welling, M., editors, *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 5 of *Proceedings of Machine Learning Research*, pages 73–80, Clearwater Beach, Florida, USA. PMLR.
- Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480.
- Cogley, T. and Sargent, T. J. (2005). Drift and volatilities: Monetary policies and outcomes in the post WWII U.S. *Review of Economic Dynamics*, 8(2):262–302.
- Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.
- George, E. and McCulloch, R. (1993). Variable selection via Gibbs sampling. *Journal of The American Statistical Association*, 88:881–889.
- George, E. I. and Ročková, V. (2020). Comment: Regularization via bayesian penalty mixing. *Technometrics*, 62(4):438–442.
- Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer, 2 edition.
- Hauzenberger, N., Huber, F., and Koop, G. (2024). Dynamic shrinkage priors for large time-varying parameter regressions using scalable markov chain monte carlo methods. *Studies in Nonlinear Dynamics and Econometrics*, 28(2):201–225.
- Hoerl, A. E. and Kennard, R. W. (2000). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 42(1):80–86.
- Hsiang, T. C. (1975). A bayesian view on ridge regression. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 24(4):267–268.
- James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 361–379. University of California Press.
- Johndrow, J. E., Orenstein, P., and Bhattacharya, A. (2018). Bayes shrinkage at gwas scale: Convergence and approximation theory of a scalable mcmc algorithm for the horseshoe prior.
- Koop, G. and Korobilis, D. (2023). Bayesian dynamic variable selection in high dimensions. *International Economic Review*, 64(1):1047–1074.
- Nakajima, S., Watanabe, K., and Sugiyama, M. (2019). *Variational Bayesian Learning Theory*. Cambridge University Press.
- Park, T. and Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686.
- Ročková, V. and George, E. I. (2014). EMVS: The EM approach to Bayesian variable selection. *Journal of the American Statistical Association*, 109(506):828–846.
- Ročková, V. and George, E. I. (2018). The Spike-and-Slab LASSO. *Journal of the American Statistical Association*, 113(521):431–444.

- Ročková, V. and McAlinn, K. (2021). Dynamic variable selection with spike-and-slab process priors. *Bayesian Analysis*, 16:233–269.
- Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 197–206. University of California Press.
- Stock, J. H. and Watson, M. W. (2007). Why Has U.S. Inflation Become Harder to Forecast? *Journal of Money, Credit and Banking*, 39(s1):3–33.
- Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society Series B Methodological*, 58:267–288.
- van Erp, S., Oberski, D. L., and Mulder, J. (2018). Shrinkage priors for Bayesian penalized regression. OSF Preprints cg8fq, Center for Open Science.
- Yüzbaşı, B., Arashi, M., and Akdeniz, F. (2021). Penalized regression via the restricted bridge estimator. *Soft Computing*, 25:8401–8416.
- Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320.

A. Software and implementation details for static selection examples

In our analysis, we employed the `glmnet` package (Friedman et al., 2010) to fit Ridge, Lasso, and Elastic Net models. Specifically, the Lasso model was run with `alpha = 1` and `nlambda = 50`, producing a sequence of penalty values $\{\lambda_{\text{lasso}}\}$ over which the coefficients $\hat{\beta}_j$ were estimated. The Ridge model used `alpha = 0` and `nlambda = 50`, yielding $\{\lambda_{\text{ridge}}\}$ and enforcing continuous shrinkage of all coefficients. The Elastic Net combined both ℓ_1 and ℓ_2 penalties by setting `alpha = 0.5` and `nlambda = 50`, thereby balancing sparsity (Lasso) and ridge-type regularization across $\{\lambda_{\text{elnet}}\}$.

For Bayesian spike-and-slab approaches, we utilized in the R software the `EMVS` package (Ročková and George, 2014) with $\{v_0\}$ specified on a logarithmic grid between 10^{-4} and 1 (50 points, `v0 = v0_grid`) and a slab variance of `v1 = 50`. We set `type = "fixed"` to fix the prior inclusion probability θ and `direction = "backward"` to control the EM trajectory. Additionally, we fitted a Spike-and-Slab LASSO via the `SSLASSO` package (Ročková and George, 2018), specifying `penalty = "adaptive"`, `variance = "fixed"`, `lambda1 = 0.1`, and a spike grid `lambda0 = seq(0.1, 50, length.out=50)` with `nlambda = 50` to trace the coefficient paths $\hat{\beta}_j(\lambda_0)$.