Sparse Autoregressive Processes for Dynamic Variable Selection

Veronika Ročková*

December 30, 2016

Abstract

We consider the problem of dynamic variable selection in time series regression models, where the set of active predictors is allowed to evolve over time. To capture time-varying variable selection uncertainty, we introduce new dynamic shrinkage priors for the time series of regression coefficients. These priors are characterized by two main ingredients: smooth parameter evolutions as well as intermittent zeroes for modeling predictive breaks. This is achieved with a multiple shrinkage formulation by switching between two shrinkage targets: either zero or the vicinity of the previous value. More formally, our proposed Autoregressive Spike-and-Slab Process (ASSP) priors are constructed as mixtures of two processes: a spike process for the irrelevant coefficients and a slab autoregressive process for the active coefficients. The mixing weights are themselves time-varying and depend on a lagged value of the series. A remarkable feature of ASSP is that their stationary distribution is fully known, where the marginals are characterized by benchmark spike-and-slab priors. This property guarantees marginal stability and equilibrium between selection and smoothing. The practical appeal is the availability fast MAP estimation algorithms, a strategy pursued here. By turning our priors into penalty functions, we formalize the notion of dynamic shrinkage, borrowing adaptively from both future and past information. We characterize dynamic selection thresholds for MAP smoothing and implement a one-step-late EM algorithm for efficient calculations. Illustrations show that ASSP succeed far better at finding signal relative to classical autoregressive state-space models, when many redundant predictors are present.

Keywords Autoregressive mixture processes, Dynamic sparsity; MAP smoothing; Stationarity

*Veronika.Rockova@ChicagoBooth.edu

Veronika Ročková is assistant professor in Econometrics and Statistics and James S. Kemper Foundation Faculty Scholar at the Booth School of Business of the University of Chicago.
1 Dynamic Sparsity

For dynamic linear modeling with many potential predictors, the assumption of a static generative model with a fixed subset of regressors (albeit with time-varying regressor effects) may be misleadingly restrictive. By obscuring variable selection uncertainty over time, confinement to a single inferential model may lead to poorer predictive performance, especially when the actual effective subset at each time is sparse. The potential for dynamic model selection techniques in time series modeling has been recognized (Frühwirth-Schnatter and Wagner, 2010; Groen et al., 2009; Nakajima and West, 2013; Kalli and Griffin, 2014; Chan et al., 2012). A prominent practical example is inflation forecasting, where different sets of predictors may predict well in recessions but not in expansions (Groen et al., 2009; Koop and Korobilis, 2012; Kalli and Griffin, 2014). Motivated by such contexts, we develop a new dynamic shrinkage approach for time series models that exploits time-varying predictive subset sparsity, when it in fact exists.

We present our approach in the context of dynamic linear models (DLM) (West and Harrison, 1999) that link a scalar response $y_t$ at time $t$ to a set of $p$ known regressors $x_t = (x_{t1}, \ldots, x_{tp})'$ through the relation

$$y_t = x_t' \beta_t^0 + \varepsilon_t, \quad t = 1, \ldots, T,$$

where $\beta_t^0 = (\beta_{t1}^0, \ldots, \beta_{tp}^0)'$ is a time-varying vector of regression coefficients and where the innovations $\varepsilon_t$ have arrived from $\mathcal{N}(0, \sigma_t^2)$ with $\sigma_t^2 = 1$. The challenge of estimating the $T \times p$ coefficients in (1.1), with merely $T$ observations, is typically made feasible with a smoothness inducing state-space model that treats $\{\beta_t^0\}_{t=1}^T$ as realizations from a (vector autoregressive) stochastic process. Nevertheless, any regression model with a large number of potential predictors will still be vulnerable to overfitting. This phenomenon is perhaps even more pronounced here, where the regression coefficients are forced to be dynamically intertwined to borrow strength. The major concern is that overfitted coefficient evolutions disguise true underlying dynamics and provide misleading representations with poor out-of-sample predictive performance. With many potentially irrelevant predictors, seeking sparsity is a natural remedy against the loss of statistical efficiency from such overfitting.

We shall assume that $p$ is potentially very large, where possibly only a small portion of predictors is relevant for the outcome at any given time. Besides time-varying regressor effects, we thus adopt the point of view that the regressors are allowed to enter and leave the model as time progresses,
rendering the subset selection problem ultimately dynamic. This anticipation can be reflected by the following sparsity manifestations in the matrix of regression coefficients $B_{p \times T}^0 = [\beta_0^1, \ldots, \beta_0^T]$:

(a) horizontal sparsity, where each individual time series $\{\beta_{0j}^T\}_{t=1}^T$ (for $j = 1, \ldots, p$) allows for intermittent zeroes for when $j^{th}$ predictor is not a persisting predictor at all times, (b) vertical sparsity, where only a subset of coefficients $\beta_0^t = (\beta_{01}^t, \ldots, \beta_{0p}^t)'$ (for $t = 1, \ldots, T$) will be active at the $t^{th}$ snapshot in time.

We are certainly not alone in proposing dynamic regularization for time-varying subset selection and shrinkage. This problem has been addressed in the literature by multiple authors including Groen et al. (2009); Belmonte et al. (2013); Koop and Korobilis (2012); Kalli and Griffin (2014); Nakajima and West (2013). Other related work includes shrinkage approaches towards static coefficients in time-varying models (Frühwirth-Schnatter and Wagner, 2010; Bitto and Frühwirth-Schnatter, 2016; Lopes et al., 2016). We approach the dynamic sparsity problem through the Bayesian variable selection lens and develop it further for time-domain linear models. Namely, we assume the traditional spike-and-slab setup by assigning each regression coefficient $\beta_{ij}$ a mixture prior underpinned by a binary latent indicator $\gamma_{ij}$, which flags it as being either active or inert. The spike-and-slab literature on static regression models is too voluminous to review exhaustively. Some traditional and more recent works include (Carlin and Chib, 1995; Clyde et al., 1994; George and McCulloch, 1993, 1997; Mitchell and Beauchamp, 1988; Ročková and George, 2014). The literature on time-domain spike-and-slab approaches, however, is far more sparse (George et al., 2008; Frühwirth-Schnatter and Wagner, 2010; Nakajima and West, 2013; Groen et al., 2009). Narrowing this gap, this work proposes several new dynamic extensions of popular spike-and-slab priors (George and McCulloch, 1993; Ročková and George, 2016) for time-domain models.

We should like to draw particular attention to the latent threshold process of Nakajima and West (2013), which served as a springboard for our approach. This latent threshold approach has a spike-and-slab flavor in the sense that it can be regarded as a regime switching scheme for either shrinking coefficients exactly to zero or for leaving them alone on their autoregressive path:

$$\beta_{ij} = b_{ij} \gamma_{ij}, \quad \text{where} \quad \gamma_{ij} = I(\|b_{ij}\| > d_j),$$

$$b_{ij} = \phi_{ij} + \phi_{1j}(b_{i-1j} - \phi_{0j}) + \varepsilon_t, \quad |\phi_{ij}| < 1, \quad \varepsilon_t \overset{iid}{\sim} N(0, \lambda_1).$$

The model assumes a latent stationary autoregressive $AR(1)$ process $\{b_{ij}\}_{t=1}^T$, giving rise to the
actual coefficients $\{\beta_{tj}\}_{t=1}^{T}$ only when it meanders away from a basin around zero $[-d_j, d_j]$. This process can be regarded as a dynamic extension of point-mass mixture priors that exhibit exact zeros (Mitchell and Beauchamp, 1988). However, the aspect of thresholding small nonzero values (with a static selection threshold $d_j$) aligns closely with the notion of practical significance, originally associated with continuous spike-and-slab priors (George and McCulloch, 1993). Recently, there has been a resurrection of interest in continuous spike-and-slab variants due to their amenability to fast optimization (Ročková and George, 2014, 2016). The latent threshold approach has, so far, relied on rather laborious MCMC implementations, for which optimization variants are far from obvious. In this work, we propose new dynamic continuous spike-and-slab alternatives which have appeal not only from a methodological but also a practical (optimization) viewpoint.

The main thrust of this work is to introduce the Autoregressive Spike-and-Slab Process (ASSP) priors, a new class of time series priors which induce either smoothness or shrinkage towards zero. These processes are formed as mixtures of two stationary time series: one for the active and one for the negligible coefficients. The ASSP priors pertain closely to the broader framework of mixture autoregressive (MAR) processes with a given lag, where the mixing weights are allowed to depend on time. Despite the reported success of MAR processes (and variants thereof) for modeling non-linear time series (Wong and Li, 2000, 2001; Kalliovirta et al., 2015; Wood et al., 2011), their potential as dynamic sparsity inducing priors has been unexplored. Here, we harvest this potential within a dynamic variable selection framework. The remarkable feature of the ASSP priors, which sets it apart from the latent threshold model, is that it yields benchmark continuous spike-and-slab priors (such as the Spike-and-Slab LASSO of Ročková (2015)) as its marginal stationary distribution. This property guarantees marginal stability and equilibrium in the selection/shrinkage dynamics.

By turning our time-domain priors into penalty constructs, we formalize the notion of prospective and retrospective shrinkage through doubly adaptive shrinkage terms that pull together past, current and future information. We introduce asymmetric dynamic thresholding rules, extensions of existing rules for static symmetric regularizers (Fan and Li, 2001; Antoniadis and Fan, 2001), to characterize the behavior of joint posterior modes for MAP smoothing. For calculation, we implement a one-step-late EM algorithm of (Green, 1990), that capitalizes on fast closed-form one-site updates. Our dynamic penalties can be regarded as natural extensions of the spike-and-slab penalty functions introduced by Ročková (2015) and further developed by Ročková and George (2016) and Ročková and George (2015). The ASSP priors here are deployed as a fast MAP smoothing catalyst rather than
a vehicle for a full-blown MCMC analysis (as in Nakajima and West (2013)). The key distinguishing feature of this deployment is that ASSP priors attain sparsity through sparse posterior modes rather than auxiliary latent thresholding.

The paper is structured as follows: Section 2 introduces the ASSP processes and discusses some of their theoretical properties. Section 3 develops the penalized likelihood perspective, introducing the prospective and retrospective shrinkage terms and characterizing selection thresholds for MAP estimates. Section 5 then develops the one-step-late EM algorithm for MAP smoothing. Section 6 illustrates the MAP smoothing deployment of ASSP on simulated examples and Section 7 concludes with a discussion.

2 Autoregressive Spike-and-Slab Processes Priors

In this section, we introduce the class of Autoregressive Spike-and-Slab Process (ASSP) priors, dynamic elaborations of spike-and-slab priors for time-domain linear models. We will assume that the $p$ time series $\{\beta_j\}_{t=1}^T$ (for $j = 1, \ldots, p$) in (1.1) follow independent and identical ASSP priors. Thus, for notational simplicity, let us focus for the moment on a single series and suppress the subscript $j$.

We begin with the conditional formulation of the ASSP priors. Given a binary indicator $\gamma_t \in \{0, 1\}$, which encodes the spike/slab membership at time $t$, and a lagged value $\beta_{t-1}$, we assume that $\beta_t$ arises from a mixture of the form

$$\pi(\beta_t | \gamma_t, \beta_{t-1}) = (1 - \gamma_t) \psi_0(\beta_t | \lambda_0) + \gamma_t \psi_1(\beta_t | \mu_t, \lambda_1),$$

(2.4)

where

$$\mu_t = \phi_0 + \phi_1 (\beta_{t-1} - \phi_0) \quad \text{with} \quad |\phi_1| < 1$$

(2.5)

and

$$P(\gamma_t = 1 | \beta_{t-1}) = \theta_t.$$  

(2.6)

The spike $\psi_0(\beta | \lambda_0)$ can be in principle any density sufficiently peaked around zero. A purposeful choice for us will be the Laplace density $\psi_0(\beta | \lambda_0) = \frac{\lambda_0}{2} e^{-|\beta|\lambda_0}$ (with a relatively large penalty parameter $\lambda_0 > 0$) due to its ability to threshold via sparse posterior modes. Regarding the slab distribution $\psi_1(\beta_t | \mu_t, \lambda_1)$, we require that it be moderately peaked around its mean $\mu_t$, where the amount of spread is regulated by $\lambda_1 > 0$. Throughout this work, we will choose $\psi_1(\beta_t | \mu_t, \lambda_1)$ to be the...
Gaussian density with mean $\mu_t$ and variance $\lambda_1$, due to its ability to smooth over past/future values. As detailed in Section 2.2, other reasonable choices of $\psi_0(\cdot)$ and $\psi_1(\cdot)$ may be considered. Note also that our framework can be naturally extended by replacing (2.5) with higher-order autoregressive polynomials where $\mu_t$ may also depend on values older than $\beta_{t-1}$.

The mixture formulation (2.4) generalizes existing continuous spike-and-slab priors (George and McCulloch, 1993; Ishwaran and Rao, 2005; Ročková and George, 2016) in two important ways. First, rather than centering the slab around zero, an ASSP prior anchors it around a model for the time-varying mean $\mu_t$, a new distinctive feature. The non-central mean is defined as an autoregressive lag polynomial of the first order with fixed hyper-parameters $(\phi_0, \phi_1)$. Albeit estimable subject to stationary restrictions, these parameters will be treated as if they were known. It is illuminating to regard the conditional prior (2.4) as a “multiple shrinkage” prior (George, 1986b,a) with two shrinkage targets: (1) zero (for the gravitational pull of the spike), and (2) $\mu_t$ (for the gravitational pull of the slab). It is also worthwhile to note that the spike distribution $\psi_0(\beta_t | \lambda_0)$ does not depend on $\beta_{t-1}$, only the slab does. The ASSP formulation thus induces segregation into two groups of regression coefficients, where only the active ones are assumed to walk on an autoregressive path.

The second important and distinctive generalization is implicitly hidden in a further hierarchical formulation for the mixing weights $\theta_t$ in (2.6) which casts them as a smoothly evolving process (as will be seen in Section 2.0.1 below). Before turning to this formulation, however, it is useful to point out the time-varying conditional moments of ASSP, given the past value $\beta_{t-1}$ after margining over $\gamma_t$:

$$E[\beta_t | \beta_{t-1}] = \theta_t[\phi_0 + \phi_1(\beta_{t-1} - \phi_0)]$$

and

$$\text{Var}[\beta_t | \beta_{t-1}] = (1 - \theta_t) \frac{2}{\lambda^2_0} + \theta_t\lambda_1 + (1 - \theta_t)\theta_t\mu^2_t.$$

Finally, we note that the conditional form (2.4)-(2.6) is a mixture of two stationary processes. Under the spike distribution, the series $\{\beta_t\}_{t=1}^T$ is trivially stationary, iid with a marginal density $\psi_0(\beta | \lambda_0)$. Whereas under the slab distribution, $\{\beta_t\}_{t=1}^T$ follow a stationary Gaussian $AR(1)$ process

$$\beta_t = \phi_0 + \phi_1(\beta_{t-1} - \phi_0) + e_t, \quad |\phi_1| < 1, \quad e_t \sim \mathcal{N}(0, \lambda_1), \quad (2.7)$$
whose the stationary distribution is characterized by univariate marginals

$$\psi_{ST}^{ST}(\beta_t | \lambda_1, \phi_0, \phi_1) \equiv \psi_1 \left( \beta_t \left| \phi_0, \frac{\lambda_1}{1-\phi_1^2} \right. \right),$$

(2.8)

the Gaussian density with mean $\phi_0$ and variance $\frac{\lambda_1}{1-\phi_1^2}$. The availability of this tractable stationary distribution (2.8) is a major appeal of the conditional Gaussian slab distribution. However, the ASSP construction is not confined to the Gaussian slab (Laplace spike). We elaborate on alternative choices in Section 2.2.

2.0.1 □ Evolving Inclusion Probabilities A very appealing feature of ASSP priors that makes them suitable for dynamic subset selection is the opportunity they afford for obtaining smooth spike/slab memberships. Recall that the binary indicators in (2.6) determine which of the spike or slab regimes is switched on at time $t$, where $P(\gamma_t = 1 | \beta_{t-1}) = \theta_t$. It is desirable that the sequence of slab probabilities $\{\theta_t\}_{t=1}^T$ evolves smoothly over time, allowing for changes in variable importance as time progresses, and at the same time, avoiding erratic regime switching. A viable strategy might be to treat $\{\theta_t\}_{t=1}^T$ as random with an autoregressive process that relates $\theta_t$ to the previous value $\theta_{t-1}$, for example with a conditional beta autoregressive process (Casarin et al., 2012) or a marginal beta autoregressive process (McKensie, 1985). Because the series $\{\theta_t\}_{t=1}^T$ is a key driver of the sparsity pattern, it is important that it be (marginally) stable and that it reflects all relevant information, including not only the previous value $\theta_{t-1}$, but also the previous value $\beta_{t-1}$. The beta autoregressive constructions mentioned above can be trivially modified to include $\beta_{t-1}$. However, characterizing stationarity becomes a far more delicate task due to the interplay between the two autoregressive processes (on $\{\beta_t\}_{t=1}^T$ and $\{\theta_t\}_{t=1}^T$).

Although the $\{\beta_t\}_{t=1}^T$ process will be stationary under each of the spike and slab distributions separately, it is not immediately obvious that it will be stationary under the spike-and-slab mixture where the $\beta_t$’s can transition between these distributions and where $\theta_t$ depends on $\beta_{t-1}$. However, with a suitable formulation for the $\{\theta_t\}_{t=1}^T$ sequence, the stability and coherence of the ASSP can be maintained. Under this formulation (introduced below), the $\{\beta_t\}_{t=1}^T$ process will not only be stationary, but will have spike-and-slab marginals. Such a $\theta_t$ sequence is obtained with a deterministic transition function of the lagged $\beta_t$ values $\theta_t = \theta(\beta_{t-1})$. For our formulation, we introduce a marginal importance weight $0 < \Theta < 1$, a scalar parameter which controls the overall balance between the spike and the slab distributions. Given $(\Theta, \lambda_0, \lambda_1, \phi_0, \phi_1)$, the conditional inclusion probability $\theta_t$ (or
a transition function $\theta(\beta_{t-1})$ is defined as

$$
\theta_t \equiv \theta(\beta_{t-1}) = \frac{\Theta \psi^{ST}_1 (\beta_{t-1} | \lambda_1, \phi_0, \phi_1)}{\Theta \psi^{ST}_1 (\beta_{t-1} | \lambda_1, \phi_0, \phi_1) + (1 - \Theta) \psi_0 (\beta_{t-1} | \lambda_0)}.
$$

Before turning to stationarity properties of the full ASSP priors, we pause a bit to appreciate the probabilistic meaning of (2.9). The conditional mixing weight $\theta_t$ can be interpreted as the posterior probability of classifying the past coefficient $\beta_{t-1}$ as arriving from the stationary slab distribution as opposed to the (stationary) spike distribution. This interpretation reveals how the weights \{\theta_t\}_{t=1}^T proliferate parsimony throughout the process \{\beta_t\}_{t=1}^T. Suppose that $|\beta_{t-1}|$ was large, then $\theta(\beta_{t-1})$ will be large as well, signaling that the current observation $\beta_t$ is more likely to be in the slab. The contrary occurs when $|\beta_{t-1}|$ is small, where $\beta_t$ is discouraged from the slab by a smaller inclusion weight $\theta(\beta_{t-1})$. It is tempting to regard $\Theta$ as the marginal proportion of nonzero coefficients. However, such an interpretation is misleading because sparsity levels are ultimately determined by the $\theta_t$ sequence which is influenced by the component stationary distributions $\psi_0(\cdot)$ and $\psi^{ST}_1(\cdot)$, in particular by the amount of their overlap around zero. Such an interpretation is thus obscured for the continuous spike-and-slab mixtures considered here, where more caution is needed for calibration (Ročková, 2015). We will come back to this issue in Section 3. Before proceeding, let us also note that the weights in (2.9) are different from the conditional probabilities for classifying $\beta_{t-1}$ as arising from the conditional slab in (2.4). These weights will be introduced later in Section 3.

A related deterministic transition function was proposed by Nakajima and West (2013) (see (1.2) and (1.3)). Our formulation, however, yields a process \{\beta_t\}_{t=1}^T with a completely characterized marginal stationary distribution.

Now that we have elaborated on all the layers of the hierarchical model, we are ready to formally define the Autoregressive Spike-and-Slab Process.

**Definition 1.** Equations (2.4), (2.5), (2.6) and (2.9) define an Autoregressive Spike-and-Slab Process (ASSP) with parameters $(\Theta, \lambda_0, \lambda_1, \phi_0, \phi_1)$. We will write

$$
\{\beta_t\}_{t=1}^T \sim \text{ASSP}(\Theta, \lambda_0, \lambda_1, \phi_0, \phi_1).
$$

The ASSP process relates to the Gaussian mixture of autoregressive (GMAR) processes of Kalliovirta et al. (2015) with the following differences: ASSP assumes two mixture components and,
more importantly, more general spike/slab distributions (as elaborated further in Section 2.2). Now, we take a closer look at the some of the properties of the ASSP process.

2.1 Stationarity

The ASSP construction has a strong appeal from a methodological point of view in the sense that its marginal probabilistic structure is fully known and can be described with mixtures of familiar distributions. This property is rarely available with conditionally defined non-Gaussian time series models, where not much is known about the stationary distribution beyond just the mere fact that it exists. The ASSP process, on the other hand, guarantees well behaved stable marginals that can be described through benchmark spike-and-slab priors.

Being inherently a mixture of stationary processes, the ASSP process ought to be stationary. In the context of MAR models, Wong and Li (2000) describe stationarity restrictions on the autoregressive polynomial parameters as well as mixing weights that are not time varying. If the mixing weights \( \theta_t \) were fixed, the stationarity would be inherited from the slab Gaussian AR(1) process when \( |\phi_1| < 1 \). Going further, Kalliovirta et al. (2015) characterize the stationary distribution of the GMAR process with time varying weights, which aligns closely with the ASSP process. The following theorem generalizes their result.

**Theorem 1.** Assume \( \{\beta_t\}_{t=1}^T \sim \text{ASSP}(\Theta, \lambda_0, \lambda_1, \phi_0, \phi_1) \) with \( |\phi_1| < 1 \). Then \( \{\beta_t\}_{t=1}^T \) has a stationary distribution characterized by the following univariate marginal distributions:

\[
\pi^{ST}(\beta | \Theta, \lambda_0, \lambda_1, \phi_0, \phi_1) = \Theta \psi^{ST}_{1}(\beta | \lambda_1, \phi_0, \phi_1) + (1 - \Theta) \psi_0(\beta | \lambda_0),
\]

(2.10)

where \( \psi^{ST}_{1}(\beta | \lambda_1, \phi_0, \phi_1) \) is the stationary slab distribution (2.8).

**Proof.** We assume an initial condition \( \beta_{t=0} \sim \pi^{ST}(\beta_0 | \Theta, \lambda_0, \lambda_1, \phi_0, \phi_1) \). Recall that the conditional density of \( \beta_1 \) given \( \beta_0 \) can be written as

\[
\pi(\beta_1 | \beta_0) = (1 - \theta_1) \psi_0(\beta_1 | \lambda_0) + \theta_1 \psi_1(\beta_1 | \mu_1, \lambda_1).
\]

(2.11)

From the definition of \( \theta_1 \) in (2.9), we can write the joint distribution as

\[
\pi(\beta_1, \beta_0) = \Theta \psi^{ST}_{1}(\beta_0 | \lambda_1, \phi_0, \phi_1) \psi_1(\beta_1 | \mu_1, \lambda_1) + (1 - \Theta) \psi_0(\beta_0 | \lambda_0) \psi_0(\beta_1 | \lambda_0).
\]
Integrating $\pi(\beta_1, \beta_0)$ with respect to $\beta_0$, we obtain

$$
\pi(\beta_1) = \int \pi(\beta_1, \beta_0) d\beta_0 = \Theta \left[ \int_{\beta_0} \psi_1(\beta_1 | \mu_1, \lambda_1) \psi_1 \left( \beta_0 | \phi_0, \frac{\lambda_1}{1 - \phi_1^2} \right) d\beta_0 \right] + (1 - \Theta) \psi_0(\beta_1 | \lambda_0).
$$

Theorem 1 describes the very elegant property of ASSP that the univariate marginals of this mixture process are $\Theta$-weighted mixtures of marginals. It also suggests a more general recipe for mixing multiple stationary processes through the construction of mixing weights (2.9). In the next section, we provide examples of such mixing with Laplace slab distributions.

### 2.2 Other Spike and Slab Densities

Rather than shrinking to the vicinity of the past value, one might like to entertain the possibility of shrinking exactly to the past value. Such property would be appreciated, for instance, in dynamic sparse portfolio allocation models to mitigate transaction costs associated with negligible shifts in the portfolio weights (Kaoru and West, 2016; Brodie et al., 2009; Jagannathan and Ma, 2003; Puelz et al., 2016). One way of attaining the desired effect would be replacing the Gaussian slab $\psi_1(\cdot)$ in (2.4) with a Laplace distribution centered at $\mu_t$. This conditional construction, however, does not imply the Laplace distribution marginally. The univariate marginals are defined through the characteristic function given in (2.7) of Anděl (1983). The lack of availability of the marginal density thwarts the specification of transition weights (2.9) needed within our ASSP framework. There are, however, avenues for constructing an autoregressive process with Laplace marginals, e.g. through the normal-gamma-autoregressive (NGAR) process by Kalli and Griffin (2014). We define the following Laplace autoregressive (LAR) process as a special case.

**Definition 2.** We define the Laplace autoregressive (LAR) process by

$$
\beta_t = \sqrt{\psi_t / \psi_{t-1}} \phi_{t-1} + \eta_t, \quad \eta_t \sim \mathcal{N}(0, (1 - \phi_1^2) \psi_t),
$$

where $\{\psi_t\}_{t=1}^T$ follow an exponential autoregressive process specified through $\psi_t | \kappa_{t-1} \sim \text{Gamma}(1 + \kappa_{t-1}, \lambda_1^2/[2(1 - \rho)])$ and $\kappa_{t-1} | \psi_{t-1} \sim \text{Poisson} \left( \frac{\rho}{2(1 - \rho)} \lambda_1^2 \psi_{t-1} \right)$ with a marginal distribution $\text{Exp}(\lambda_1^2/2)$.

The LAR process exploits the scale-normal-mixture representation of the Laplace distribution, yielding Laplace marginals $\beta_t \sim \tilde{\psi}_{ST}(\beta_t | \lambda_1) \equiv \text{Laplace}(\lambda_1)$. This coherence property can be leveraged...
within our ASSP framework as follows. If we replace the slab Gaussian AR(1) process in (2.4) with the LAR process and deploy \( \tilde{\psi}_{ST}(\beta_t | \lambda_1) \) instead of \( \psi_{ST}(\beta_t | \lambda_1) \) in (2.9), we obtain a Laplace ASSP variant with the Spike-and-Slab LASSO prior of Ročková (2015) as its marginal distribution (according to Theorem 1).

It is worth pointing out an alternative autoregressive construction with Laplace marginals proposed by Anděl (1983). He considered the following AR(1) scheme

\[
\beta_t = \begin{cases} 
\phi_1 \beta_{t-1} & \text{with probability } \phi_1^2, \\
\phi_1 \beta_{t-1} + \eta_t & \text{with probability } 1 - \phi_1^2, \end{cases} \quad \text{where } \eta_t \sim \text{Laplace}(\lambda_1). \tag{2.12}
\]

The innovations in (2.12) come from a mixture of a point mass at zero, providing an opportunity to settle at the previous value, and a Laplace distribution. Again, deploying this process in the slab, we obtain the Spike-and-Slab LASSO marginal distribution. Despite attractive methodologically, these Laplace extensions are ultimately less appetizing for implementation. Throughout the rest of the paper, we focus on the Gaussian AR(1) slab process.

Before proceeding let us note that the spike distribution \( \psi_0(\beta | \lambda_0) \) can be replaced by any (continuous) density without disturbing the validity of Theorem 1. A Gaussian spike, for instance, would impose no new computational challenges due to its conditional conjugacy. However, additional thresholding would be required to obtain sparse posterior modes. In the sequel, we focus on the Laplace spike due to its automatic thresholding property.

### 3 Dynamic Spike-and-Slab Penalty

Spike-and-slab priors give rise to self-adaptive penalty functions for MAP estimation, as detailed in Ročková (2015) and Ročková and George (2016). Here we introduce new penalty constructs for dynamic shrinkage implied by the ASSP priors.

**Definition 3.** For a given set of parameters \( (\Theta, \lambda_0, \lambda_1, \phi_0, \phi_1) \), we define a prospective penalty function implied by (2.4) and (2.9) as follows:

\[
\text{pen} (\beta | \beta_{t-1}) = \log \left[ (1 - \theta_t) \psi_0(\beta | \lambda_0) + \theta_t \psi_1(\beta | \mu_t, \lambda_1) \right]. \tag{3.13}
\]

Similarly, we define a retrospective penalty \( \text{pen}(\beta_{t+1} | \beta) \) as a function of the second argument \( \beta \) in...
Figure 1: Plots of the prospective penalty function

\[ (3.13) \]

\[
Pen(\beta | \beta_{t-1}, \beta_{t+1}) = pen(\beta | \beta_{t-1}) + pen(\beta_{t+1} | \beta) + C,
\]

where \( C \equiv -Pen(0 | \beta_{t-1}, \beta_{t+1}) \) is a norming constant such that \( Pen(0 | \beta_{t-1}, \beta_{t+1}) = 0 \).

**Remark 1.** Note that the dependence on the previous value \( \beta_{t-1} \) in \( pen(\beta | \beta_{t-1}) \) is hidden in \( \theta_t \) and \( \mu_t \). Throughout the paper, we will write \( \partial \theta_t/\partial \beta_{t-1} \) and \( \partial \mu_t/\partial \beta_{t-1} \) without reminding ourselves of this implicit relationship.

To gain some insights about the prospective penalty (3.13), it is helpful to point out one special case when \( \mu_t = 0 \). Then both the spike and the slab are centered at zero, yielding a penalty reminiscent of the Spike-and-Slab LASSO by Ročková (2015). However, instead of mixing two LASSO penalties, here we have an adaptive variant of the elastic net (Zou and Hastie, 2005). With \( \mu_t \neq 0 \), the prospective penalty is no longer symmetric around zero (and not guaranteed to be concave). Figure 1 portrays the prospective penalty for two choices of \( \beta_{t-1} \) and two sets of tuning parameters \( \phi_1, \lambda_1, \lambda_0 \) and \( \Theta \) (assuming \( \phi_0 = 0 \)). Because the conditional transfer equation (2.4) is a mixture, \( pen(\beta | \beta_{t-1}) \) is apt to be multimodal. Figure 1(a) shows an obvious peak at zero (due to the Laplace spike), but also a peak around \( \mu_t = 0.9 \times \beta_{t-1} \), prioritizing values in the close vicinity of the previous value (due to the non-central slab). From the implementation viewpoint, however, it is more desirable that the penalty be uni-modal, reflecting the size of the previous coefficient without ambiguity by suppressing one of the peaks. Such behavior is illustrated in Figure 1(b) and Figure 1(c), where the penalty flexibly adapts to \( |\beta_{t-1}| \) by promoting either zero or a value close to \( \beta_{t-1} \). This effect is achieved...
with a relatively large stationary slab variance $\lambda_1/(1 - \phi_1^2) = 10$, a mild Laplace peak $\lambda_0 = 1$ and the marginal importance weight $\Theta = 0.9$. Smaller values $\Theta$ would provide an overwhelming support for the zero mode. The parameter $\Theta$ thus should not be regarded as a proportion of active coefficients (as is customary with point-mass mixtures) but rather an interpretation-free tuning parameter.

Figure 1 plots $\text{pen}(\beta | \beta_{t-1})$ prospectively as a function of $\beta$, given the previous value $\beta_{t-1}$. It is also illuminating to plot $\text{pen}(\beta_{t+1} | \beta)$ retrospectively as a function of $\beta$, given the future value $\beta_{t+1}$. Two such retrospective penalty plots are provided in Figure 2(a) and Figure 2(b). When the future value is relatively large ($\beta_{t+1} = 1.5$ in Figure 2(b)), the penalty $\text{pen}(\beta_{t+1} | \beta)$ has a peak near $\beta_{t+1}$, signaling that the value $\beta_t$ must be large too. When the future value is small ($\beta_{t+1} = 0$ in Figure 2(a)), the penalty has a peak at zero signaling that the current value $\beta_t$ must be small. Again, this balance is achieved with a relatively large stationary slab variance and a large value $\Theta$.

The behavior of the prospective and retrospective penalties is ultimately tied to the mixing weight $\theta_t \equiv \theta(\beta)$ in (2.9). It is desirable that $\theta(\beta)$ is increasing with $|\beta|$. However, Laplace tails will begin to dominate for large enough $|\beta|$, where the probability $\theta(\beta)$ will begin to drop (for $|\beta|$ greater than $\delta \equiv (\lambda_0 + \sqrt{2C/A})A$, where $A = \lambda_1/(1 - \phi_1^2)$ and $C = \log((1 - \Theta)/\Theta\lambda_0/2\sqrt{2\pi A})$). However, we can make the turning point $\delta$ large enough with larger values $\Theta$ and smaller values $\lambda_0$, as indicated in Figure 2(c).

To fully grasp the shrinkage dynamics implied by the penalty (3.14), it is useful to study the partial derivative $\partial \text{Pen}(\beta | \beta_{t-1}, \beta_{t+1})/\partial |\beta|$. This term encapsulates how much shrinkage we expect at time $t$, conditionally on $(\beta_{t-1}, \beta_{t+1})$. We will separate the term into two pieces: a prospective shrinkage effect $\lambda^*(\beta | \beta_{t-1})$, driven by the past value $\beta_{t-1}$, and a retrospective shrinkage effect $\tilde{\lambda}^*(\beta | \beta_{t+1})$.
driven by the future value $\beta_{t+1}$. More formally, we write

$$\frac{\partial \text{Pen}(\beta | \beta_{t-1}, \beta_{t+1})}{\partial |\beta|} = -\Lambda^*(\beta | \beta_{t-1}, \beta_{t+1}),$$

where

$$\Lambda^*(\beta | \beta_{t-1}, \beta_{t+1}) = \lambda^*(\beta | \beta_{t-1}) + \bar{\lambda}^*(\beta | \beta_{t+1}),$$

and

$$\lambda^*(\beta | \beta_{t-1}) = -\frac{\partial \text{pen}(\beta | \beta_{t-1})}{\partial \beta|},$$

and

$$\bar{\lambda}^*(\beta | \beta_{t+1}) = -\frac{\partial \text{pen}(\beta_{t+1} | \beta)}{\partial \beta|}.$$  

3.0.1 □ Shrinkage "from the Past" The prospective shrinkage term $\lambda^*(\beta | \beta_{t-1})$ pertains closely to Bayesian penalty mixing introduced by Ročková (2015) and Ročková and George (2016) in the sense that it can be characterized as an adaptive linear combination of individual spike and slab shrinkage terms. In particular, we can write

$$\lambda^*(\beta | \beta_{t-1}) = -p^*_t(\beta) \frac{\partial \log \psi_1(\beta | \mu_t, \lambda_1)}{\partial \beta} - [1 - p^*_t(\beta)] \frac{\partial \log \psi_0(\beta | \lambda_0)}{\partial \beta},$$

$$= p^*_t(\beta) \left( \frac{\beta - \mu_t}{\lambda_1} \right) \text{sign}(\beta) + [1 - p^*_t(\beta)] \lambda_0.$$  

where

$$p^*_t(\beta) = \frac{\theta_t \psi_1(\beta | \mu_t, \lambda_1)}{\theta_t \psi_1(\beta | \mu_t, \lambda_1) + (1 - \theta_t) \psi_0(\beta | \lambda_0)}.$$  

Two observations are in order: the first is that by writing $p^*_t(\beta) = P(\gamma_t = 1 | \beta_t = \beta, \beta_{t-1}, \theta_t)$(3.18) can be viewed as a posterior probability for classifying $\beta$ as arising from the conditional slab (versus the spike) at time $t$, given the previous value $\beta_{t-1}$. Second, these weights are very different from $\theta_t$ in (2.9), which are classifying $\beta$ as arising from the marginal slab (versus the spike). From (3.18), we can see how $p^*_t(\beta)$ hierarchically transmits information about the past value $\beta_{t-1}$ (via $\theta_t$) to determine the right shrinkage for $\beta_t$. This is achieved with a doubly-adaptive chain reaction. Namely, if the previous value $\beta_{t-1}$ was large, $\theta_t$ will be close to one signaling that the next coefficient $\beta_t$ is prone to be in the slab. Next, if $\beta_t$ is in fact large, $p^*_t(\beta_t)$ will be close to one, where the first summand in (3.16) becomes the leading term and shrinks $\beta_t$ towards $\mu_t$. If $\beta_t$ is small, however, $p^*_t(\beta_t)$ will be small as well, where the second term in (3.16) takes over to shrink $\beta_t$ towards zero. This gravitational pull is
accelerated when the previous value $\beta_{t-1}$ was negligible (zero), in which case $\theta_t$ will be even smaller, making it even more difficult for the next coefficient $\beta_t$ to escape the spike. This mechanism explains how the prospective penalty adapts to both $(\beta_{t-1}, \beta_t)$, promoting smooth forward proliferation of spike/slab allocations and coefficients.

3.0.2 □ Shrinkage “from the Future” While the prospective shrinkage term promotes smooth forward proliferation, the retrospective shrinkage term $\tilde{\lambda}^*(\beta | \beta_{t+1})$ operates backwards. We can write

$$\tilde{\lambda}^*(\beta | \beta_{t+1}) = -\frac{\partial \theta_{t+1}}{\partial \beta} \left[ \frac{p^*_{t+1}(\beta_{t+1})}{\theta_{t+1}} \left( 1 - p^*_{t+1}(\beta_{t+1}) \right) \right] - p^*_{t+1}(\beta_{t+1}) \phi_1 \text{sign}(\beta) \left[ \frac{\beta_{t+1} - \mu_{t+1}}{\lambda_1} \right],$$

(3.19)

where

$$\frac{\partial \theta_{t+1}}{\partial \beta} = \theta_{t+1}(1 - \theta_{t+1}) \left[ \lambda_0 - \text{sign}(\beta) \left( \frac{\beta - \phi_0}{\lambda_1/(1 - \phi_2^2)} \right) \right].$$

(3.20)

For simplicity, we will write $p^*_{t+1} = p^*_{t+1}(\beta_{t+1})$. Then we have

$$\tilde{\lambda}^*(\beta | \beta_{t+1}) = \left[ \lambda_0 - \text{sign}(\beta) \left( \frac{\beta - \phi_0}{\lambda_1/(1 - \phi_2^2)} \right) \right] \left[ (1 - p^*_{t+1}) \theta_{t+1} - p^*_{t+1}(1 - \theta_{t+1}) \right]$$

(3.21)

$$- p^*_{t+1} \phi_1 \text{sign}(\beta) \left[ \frac{\beta_{t+1} - \mu_{t+1}}{\lambda_1} \right].$$

(3.22)

The retrospective term synthesizes information from both $(\beta_{t+1}, \beta_t)$ to contribute to shrinkage at time $t$. When $(\beta_{t+1}, \beta_t)$ are both large, we obtain $p^*_{t+1}(\beta_{t+1})$ and $\theta_{t+1}$ that are both close to one. The shrinkage is then driven by the second summand in (3.22), forcing $\beta_t$ to be shrunk towards the future value $\beta_{t+1}$ (through $\mu_{t+1} = \phi_0 + \phi_1(\beta_t - \phi_0)$). When either $\beta_{t+1}$ or $\beta_t$ are small, shrinkage is targeted towards the stationary mean through the dominant term (3.21).

4 Characterization of the Joint Posterior Mode

Unlike the majority of previous developments, this paper views Bayesian dynamic shrinkage through optimization lens. Rather than distilling posterior samples to learn about $\beta = [\beta_1, \ldots, \beta_T]$, we focus on finding the MAP trajectory $\hat{\theta} = \arg \max \pi(\beta | y)$. MAP sequence estimation problems (for non-linear non-Gaussian dynamic models) were addressed previously with e.g. Viterbi-style algorithms (Godsill et al., 2001). Our optimization strategy is conceptually very different. The key to our approach will be drawing upon the penalized likelihood perspective developed in Section 3. Namely, we develop a new dynamic coordinate-wise strategy, building on existing developments for static high-dimensional variable selection.
To set the stage, we will first assume \( p = 1 \) and \( x_t = 1 \) for \( t = 1, \ldots, T \). Later, we expand our findings to the full-blown dynamic model (with \( p > 1 \)).

4.0.1 □ The Simplest One-dimensional Case To illustrate the functionality of the dynamic penalty from Section 3, we start by assuming \( p = 1 \) and \( x_t = 1 \) in (1.1). This simple case corresponds to a sparse normal-means model, where the means are dynamically intertwined. We begin by characterizing some basic properties of the posterior mode

\[
\tilde{\beta} = \arg \max_{\beta} \pi(\beta | y),
\]

where \( y = (y_1, \ldots, y_T)' \) arises from (1.1) and \( \beta = (\beta_1, \ldots, \beta_T)' \) is assigned the ASSP prior.

One of the attractive features of the Laplace spike in (2.4) is that \( \tilde{\beta} \) has a thresholding property. This property is revealed from necessary characterizations for each \( \tilde{\beta}_t \) (for \( t = 1, \ldots, T \)), once we condition on the rest of the directions through \((\tilde{\beta}_{t-1}, \tilde{\beta}_{t+1})\). The conditional thresholding rule can be characterized using standard arguments, as with similar existing regularizers (Zhang, 2010; Fan and Li, 2001; Antoniadis and Fan, 2001; Zhang and Zhang, 2012; Ročková and George, 2016). There is, however, one very important difference. While the typical sparsity-inducing penalty functions are symmetric, the penalty (3.14) is not, due to its dependence on the previous and future values \((\beta_{t-1}, \beta_{t+1})\). Thereby, instead of a single selection threshold, we have two selection thresholds

\[
\Delta_t^- = \sup_{\beta < 0} \left\{ \frac{\beta}{2} + \frac{\text{Pen}(\beta | \beta_{t-1}, \beta_{t+1})}{\beta} \right\} \quad \text{and} \quad \Delta_t^+ = \inf_{\beta > 0} \left\{ \frac{\beta}{2} + \frac{\text{Pen}(\beta | \beta_{t-1}, \beta_{t+1})}{\beta} \right\} \quad (4.23)
\]

to determine where to shrink the observation \( y_t \) at time \( t \).

**Lemma 1.** Let \( \tilde{\beta} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_T)' \) denote the global mode of \( \pi(\beta | y) \). Then, conditionally on \((\tilde{\beta}_{t-1}, \tilde{\beta}_{t+1})\), we have for \( 1 < t < T \)

\[
\tilde{\beta}_t = 0 \quad \text{iff} \quad \Delta_t^- < y_t < \Delta_t^+.
\]

**Proof.** We begin by noting that \( \tilde{\beta}_t \) is a maximizer in \( t^{th} \) direction while keeping \((\tilde{\beta}_{t-1}, \tilde{\beta}_{t+1})\) fixed, i.e.

\[
\tilde{\beta}_t = \arg \max_{\beta} \left\{ -\frac{1}{2}(y_t - \beta)^2 + \text{Pen}(\beta | \tilde{\beta}_{t-1}, \tilde{\beta}_{t+1}) \right\}. \quad (4.24)
\]

It turns out that \( \tilde{\beta}_t = 0 \) iff \( \beta \left( \frac{\beta}{2} + \frac{\text{Pen}(\beta | \tilde{\beta}_{t-1}, \tilde{\beta}_{t+1})}{\beta} - y_t \right) < 0, \forall \beta \in \mathbb{R} \setminus \{0\} \) (Zhang and Zhang, 2012).
The rest of the proof follows from the definition of $\Delta_t^+$ and $\Delta_t^-$ in (4.23).

**Remark 2.** While Lemma 1 assumes $1 < t < T$, the characterization applies also for $t = 1$, once we specify the initial condition $\beta_{t=0}$. The value $\beta_{t=0}$ is not assumed known and will be estimated together with all the remaining parameters (Section 5). For $t = T$, an analogous characterization exists, where the shrinkage term and the selection threshold only contain the prospective portion of the penalty.

The following necessary characterization links the behavior of $\hat{\beta}$ to the shrinkage terms characterized in Section 3.0.1 and Section 3.0.2.

**Lemma 2.** Let $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_T)'$ denote the global mode of $\pi(\beta | y)$. Then, conditionally on $(\hat{\beta}_{t-1}, \hat{\beta}_{t+1})$, we have for $1 < t < T$

$$
\hat{\beta}_t = \begin{cases} 
0 & \text{if } \Delta_t^- < y_t < \Delta_t^+ \\
[y_t - \Lambda^*(\hat{\beta}_t | \hat{\beta}_{t-1}, \hat{\beta}_{t+1})]_+ \text{sign}(y_t) & \text{if } \Delta_t^- \geq y_t \text{ or } y_t \geq \Delta_t^+, 
\end{cases}
$$

(4.25)

where $\Lambda^*(\hat{\beta}_t | \hat{\beta}_{t-1}, \hat{\beta}_{t+1})$ was defined in (3.15).

**Proof.** Conditionally on $(\hat{\beta}_{t-1}, \hat{\beta}_{t+1})$, the global mode $\hat{\beta}_t$, once nonzero, has to satisfy (4.25) from the first-order necessary condition. Combined with Lemma 1 we obtain the desired characterization. \(\square\)

Lemma 2 formally certifies that the mode exhibits both (a) sparsity and (b) smoothness (through the prospective/retrospective shrinkage terms). To further illustrate this behavior, we consider a simple simulated example. The true regression vector $\beta^0 = (\beta^0_1, \ldots, \beta^0_T)'$ is generated from the $AR(1)$ process (2.7) with $\phi_0 = 0, \phi_1 = 0.98$ and $\lambda_1 = 0.04$. To obtain a smooth series that is also sparse, all entries $\beta^0_t$ with an absolute value below 0.3 are thresholded to zero. Next, the observed series $\{y_t\}_{t=1}^{T=100}$ is generated from (1.1) with $\sigma_t^2 = 0.25$. The plot of the observed series $\{y_t\}_{t=1}^{T=100}$, as well as the true underlying signal $\{\beta^0_t\}_{t=1}^{T=100}$, is in Figure 3(a). Assuming $\beta_{t=0} = 1$, the posterior mode $\hat{\beta} = \arg \max_{\beta} \pi(\beta | y)$ is computed by cycling over one-site maximizers (4.24) with the aid of the screening rule in Lemma 1. The calculation is initialized with a zero vector and the tuning parameters are chosen as in Figure 1(b), i.e. $\phi_0 = 0, \phi_1 = 0.98, \lambda_0 = 1, \lambda_1 = 10(1 - \phi_1^2)$ and $\Theta = 0.9$. The fitted series, depicted in Figure 3(b), features the bipolar behavior described in Lemma 2: shrinkage to exact zero, when the observed signal is small, and smoothing of past/future values, when the signal
is large. In Figure 3(c), the MAP trajectory is seen to closely track the zero/nonzero regimes and it does not seem to overreact to spurious peaks in the true signal.

While the univariate model with \( p = 1 \) is certainly of independent interest, we move onto the more challenging case when \( p > 1 \). There the plot thickens substantially. With more predictors, there is a delicate balance between the multiple series of regression coefficients, where overfitting in one dimension may impair recovery in other directions.

4.0.2 □ The Case of General Regression The ideas developed in the previous section can be naturally extended to the case \( p > 1 \) and \( x_{tj} \neq 1 \). Instead of a single set of thresholds \( \Delta^+_t \) and \( \Delta^-_t \) (defined in (4.23)), we have one pair of thresholds for each time point and each of the \( p \) coordinates. Conditionally on \( (\beta_{t-1j}, \beta_{t+1j}) \), we define

\[
\Delta^+_t = \inf_{\beta>0} \left\{ \frac{\beta x^2_{tj}}{2} + \frac{\text{Pen}(\beta | \beta_{t-1j}, \beta_{t+1j})}{\beta} \right\} \quad \text{and} \quad \Delta^-_t = \sup_{\beta<0} \left\{ \frac{\beta x^2_{tj}}{2} + \frac{\text{Pen}(\beta | \beta_{t-1j}, \beta_{t+1j})}{\beta} \right\}
\]

(4.26)

We obtain analogous selection thresholds for MAP estimation.

**Lemma 3.** Let \( \hat{B} = \{\hat{\beta}_{tj}\}_{t,j=1}^T \) denote the global mode of \( \pi(B|Y) \). Let \( \hat{B}_{tj} \) denote all but the \((t,j)\)th entry in \( \hat{B} \) and by \( z_{tj} = y_t - \sum_{i \neq j} x_{ti} \hat{\beta}_{ti} \). Then, conditionally on \( \hat{B}_{tj} \), we have for \( 1 < t < T \)

\[
\hat{\beta}_{tj} = 0 \quad \text{iff} \quad \Delta^-_{tj} < x_{tj} z_{tj} < \Delta^+_{tj}.
\]
Proof. We note that $\hat{\beta}_{tj}$ is a maximizer in $(t,j)^{th}$ direction while keeping $\hat{B}_{tj}$ fixed, i.e.

$$\hat{\beta}_{tj} = \arg \max_{\beta} \left\{ -\frac{1}{2}(z_{tj} - x_{tj}\beta)^2 + Pen(\beta | \hat{\beta}_{t-1j}, \hat{\beta}_{t+1j}) \right\} .$$  \hspace{1cm} (4.27)$$

It turns out that $\hat{\beta}_{tj} = 0$ iff $\beta\left(\frac{\beta_x^2}{2} + Pen(\beta | \hat{\beta}_{t-1j}, \hat{\beta}_{t+1j}) - x_{tj} z_{tj}\right) < 0, \forall \beta \in \mathbb{R}\{0\}$. The rest of the proof follows from the definition of $\Delta^+_t$ and $\Delta^-_t$ in (4.26). \hfill \square

We immediately obtain an analog of Lemma 2.

**Lemma 4.** Let $\hat{\mathcal{B}}_t = \{\hat{\beta}_{tj}\}_{t,j=1}^{T,p}$ denote the global mode of $\pi(\mathcal{B}|Y)$. Let $\hat{\mathcal{B}}_{tj}$ denote all but the $(t,j)^{th}$ entry in $\hat{\mathcal{B}}$ and by $z_{tj} = y_t - \sum_{i \neq j} x_{ti}\hat{\beta}_{ti}$. Then $\hat{\beta}_{tj}$ satisfies the following necessary condition

$$\hat{\beta}_{tj} = \begin{cases} \frac{1}{x_{tj}} \left[ x_{tj}z_{tj} - \Lambda^*(\hat{\beta}_{tj} | \hat{\beta}_{t-1j}, \hat{\beta}_{t+1j}) \right] + \text{sign}(x_{tj}z_{tj}) & \text{if } \Delta^-_t < x_{tj}z_{tj} < \Delta^+_t \\ 0 & \text{otherwise.} \end{cases} \hspace{1cm}$$

**Proof.** The proof again follows from the first order necessary condition. \hfill \square

Lemma 4 evokes coordinate-wise optimization for obtaining the posterior mode. However, the computation of selection thresholds $(\Delta^-_t, \Delta^+_t)$ (as well as the one-site maximizers (4.27)) requires numerical optimization. The lack of availability of closed-form thresholding hampers practicality when $T$ and $p$ are even moderately large. In the next section, we propose an alternative strategy which capitalizes on closed-form thresholding rules.

## 5 One-step-late EM Algorithm

A (local) posterior mode $\hat{\mathcal{B}}$ can be obtained either directly, by cycling over one-site updates (4.27), or indirectly through an EM algorithm, a strategy pursued here. The direct algorithm consists of integrating out $\Gamma = [\gamma_1, \ldots, \gamma_T]'$ and solving a sequence of non-standard optimization problems (4.27) which necessitate numerical optimization. The EM algorithm, on the other hand, treats $\Gamma$ as missing data, obviating the need for numerical optimization by offering closed form one-site updates. We now describe this dynamic EM approach for MAP estimation under ASSP priors.

The initial vector $\beta_{t=0} = (\beta_{01}, \ldots, \beta_{0p})'$ at time $t = 0$ will be estimated together with all the
remaining coefficients $B$. We assume that $\beta_0$ comes from the hierarchical stationary distribution

$$
\pi(\beta_0 | \gamma_0) = \prod_{j=1}^{p} \left[ \gamma_0 \psi_1^{ST} (\beta_{0j} | \lambda_1, \phi_0, \phi_1) + (1 - \gamma_0) \psi_0(\beta_{0j} | \lambda_0) \right],
$$

where $\gamma_0 = (\gamma_{01}, \ldots, \gamma_{0p})'$ are independent binary indicators with $P[\gamma_{0j} = 1 | \Theta] = \Theta$ for $1 \leq j \leq p$. The goal is obtaining the joint mode $[\hat{B}, \hat{\beta}_0]$ of the functional $\pi(B, \beta_0 | Y)$. To this end, we proceed iteratively by augmenting this objective function with the missing data $[\Gamma, \gamma_0]$, as prescribed by Ročková and George (2014), and then maximizing w.r.t. $[B, \beta_0]$. An important observation, that facilitates the derivation of the algorithm, is that the prior distribution $\pi(\beta_0, B, \gamma_0, \Gamma)$ can be factorized into the following products

$$
\pi(\beta_0, B, \gamma_0, \Gamma) = \pi(\beta_0 | \gamma_0) \prod_{j=1}^{p} \prod_{t=1}^{T} \pi(\beta_{tj} | \gamma_{tj}, \beta_{t-1j}) \pi(\gamma_{tj} | \beta_{t-1j}),
$$

where $\pi(\beta_{tj} | \gamma_{tj}, \beta_{t-1j})$ and $\pi(\gamma_{tj} | \beta_{t-1j})$ are defined in (2.4) and (2.6), respectively. For simplicity (and without loss of generality), we will assume $\phi_0 = 0$ and thereby $\mu_{tj} = \phi_1 \beta_{t-1j}$. Then, we can write

$$
\log \pi(\beta_0, B, \gamma_0, \Gamma | Y) = C - \frac{1}{2} \sum_{t=1}^{T} (y_t - x_t' \beta_t)^2 - \sum_{t=1}^{T} \sum_{j=1}^{p} \left[ \gamma_{tj} \left( \beta_{tj} - \phi_1 \beta_{t-1j} \right)^2 + (1 - \gamma_{tj}) \lambda_0 | \beta_{tj} \right] + \sum_{t=1}^{T} \sum_{j=1}^{p} [\gamma_{tj} \log \theta_{tj} + (1 - \gamma_{tj}) \log(1 - \theta_{tj})] - \sum_{j=1}^{p} \left[ \gamma_{0j} \frac{\beta_0^2}{2 \lambda_1} (1 - \phi_1^2) + (1 - \gamma_{0j}) \lambda_0 | \beta_{0j} \right] + \sum_{j=1}^{p} [\gamma_{0j} \log \Theta + (1 - \gamma_{0j}) \log(1 - \Theta)].
$$

(5.29)

In the following, we will endow the parameters $[\beta_0, B]$ with a superscript $m$ to designate their most recent values at the $m^{th}$ iteration. In the E-step, we compute the expectation $E_{\gamma_0, \Gamma} \log \pi(\beta_0, B, \gamma_0, \Gamma | Y) \log \pi(\beta_0, B, \gamma_0, \Gamma | Y)$ with respect to the conditional distribution of $[\Gamma, \gamma_0]$, given $[B^{(m)}, \beta_0^{(m)}]$ and $Y$. All we have to really do is obtain $p_{tj}^* = P(\gamma_{tj} = 1 | \beta_{tj}^{(m)}, \beta_{t-1j}^{(m)}, \theta_{tj})$ from (3.18), when $t > 0$, and $p_{0j}^* \equiv \theta_{tj} \equiv \theta(\beta_{0j})$ from (2.9), and replace all the $\gamma_{tj}$’s in (5.29) with $p_{tj}^*$’s. In the M-step, we set out to maximize $E_{\gamma_0, \Gamma} \log \pi(\beta_0, B, \gamma_0, \Gamma | Y)$ w.r.t. $[\beta_0, B]$. We proceed coordinate-wise, iterating over the following single-site updates while keeping all the remaining parameters fixed. For $1 < t < T$, we have

$$
\beta_{tj}^{(m+1)} = \arg \max_{\beta} Q_{tj}(\beta),
$$
where

$$Q_{ij}(\beta) = -\frac{1}{2}(z_{ij} - x_{ij} \beta)^2 - \frac{p^*_{ij}}{2\lambda_1}(\beta - \phi_1 \beta^{(m)}_{t-j})^2 - \frac{p^*_{t+1j}}{2\lambda_1}(\beta^{(m)}_{t+1j} - \phi_1 \beta)^2$$

$$- (1 - p^*_{t+1j}) \lambda_0 |\beta| + p^*_{t+1j} \log \theta_{t+1j} + (1 - p^*_{t+1j}) \log (1 - \theta_{t+1j}),$$

and where $z_{ij} = y_t - \sum_{i \neq j} x_{ij} \beta^{(m)}_{t+1j}$. From the first-order condition, the solution $\beta^{(m+1)}_{t+1j}$, if nonzero, needs to satisfy $\partial Q_{ij}(\beta)/\partial \beta|_{\beta=\beta^{(m+1)}_{t+1j}} = 0$. To write the derivative a bit more concisely, we introduce the following notation: $Z_{ij} = x_{ij} z_{ij} + \frac{p^*_{ij} \phi_1}{\lambda_1} \beta^{(m+1)}_{t-1j} + \frac{p^*_{t+1j} \phi_1}{\lambda_1} \beta^{(m+1)}_{t+1j}$ and $W_{ij} = \left( x_{ij} + \frac{p^*_{ij} \phi_1}{\lambda_1} + \frac{p^*_{t+1j} \phi_1}{\lambda_1} \right)$. Then we can write for $\beta \neq 0$

$$\frac{\partial Q_{ij}(\beta)}{\partial \beta} = Z_{ij} - W_{ij} \beta - (1 - p^*_{t+1j}) \lambda_0 \text{sign}(\beta) + \frac{\partial \theta_{t+1j}}{\partial \beta} \left[ \frac{p^*_{t+1j}}{\theta_{t+1j} - 1 - \theta_{t+1j}} - \frac{1 - p^*_{t+1j}}{\theta_{t+1j} - 1 - \theta_{t+1j}} \right],$$

where

$$\frac{\partial \theta_{t+1j}}{\partial \beta} = \theta_{t+1j}(1 - \theta_{t+1j}) \left[ \lambda_0 \text{sign}(\beta) - \frac{\beta(1 - \phi_1^2)}{\lambda_1} \right]$$

is obtained from (3.20). Recall that $\theta_{t+1j}$, defined in (2.9), depends on $\beta_{ij}$ (denoted by $\beta$ above). This complicates the tractability of the M-step. If $\theta_{t+1j}$ was fixed, we could obtain a simple closed-form solution $\beta^{(m+1)}_{t+1j}$ through an elastic-net-like update (Zou and Hastie, 2005). We can take advantage of this fact with a one-step-late (OSL) adaptation of the EM algorithm (Green, 1990). The OSL EM algorithm bypasses intricate M-steps by evaluating the intractable portions of the penalty derivative at the most recent value, rather than the new value. We apply this trick to the last summand in (5.32). Instead of treating $\theta_{t+1j}$ as a function of $\beta$ in (5.32), we fix it at the most recent value $\beta^{(m)}_{t+1j}$. The solution for $\beta$, implied by (5.32), is then

$$\beta^{(m+1)}_{t+1j} = \frac{1}{M_{ij} + (1 - \phi_1^2)/\lambda_1 M_{ij}} [Z_{ij} - \Lambda_{ij}] \text{sign}(Z_{ij}), \quad \text{for} \quad 1 < t < T,$$

where $M_{ij} = p^*_{t+1j}(1 - \theta_{t+1j}) - \theta_{t+1j}(1 - p^*_{t+1j})$ and $\Lambda_{ij} = \lambda_0[(1 - p^*_{t+1j}) - M_{ij}]$. The update (5.33) is a thresholding rule, with a shrinkage term that reflects the size of $(\beta^{(m)}_{t-1j}, \beta^{(m)}_{t+1j}, \beta^{(m)}_{t+1j})$. The exact thresholding property is obtained from sub-differential calculus, because $Q_{ij}(\cdot)$ is not differentiable at zero (due to the Laplace spike). A very similar update is obtained also for $t = T$, where all the
terms involving $p_{t+1}^* \theta_{t+1}$ in $A_{tj}$, $W_{tj}$ and $Z_{tj}$ disappear. For $t = 0$, we have

$$
\beta_{0j}^{(m+1)} = \frac{1}{p_{tj}^* \phi_1^2 + p_{tj}^* (1 - \phi_1^2)} \left[ p_{tj}^* \beta_{1j} \phi_1 - (1 - p_{tj}^* \lambda_0 \lambda_1) \right]_+ \text{sign}(\beta_{1j}).
$$

(5.34)

The updates (5.33) and (5.34) can be either cycled-over at each M-step, or performed just once for each M-step.

6 Dynamic Variable Selection: Examples

To illustrate the ability of the ASSP priors to alleviate the curse of overfitting, we consider two simulated examples, one low-dimensional and one higher-dimensional with many redundant predictors.

6.1 Low-dimensional Case

We first illustrate our dynamic variable selection procedure on a simulated example with $T = 100$ observations generated from the model (1.1) with $p = 6$ predictors. The predictor values $x_{tj}$ are obtained independently from a standard normal distribution. Two out of the six predictors (predictors $x_{t5}$ and $x_{t6}$) never contribute to the model, where $\beta_{t5}^0 = \beta_{t6}^0 = 0$ at all times. The predictor $x_{t1}$ is a persisting predictor, where $\{\beta_{t1}\}_{t=1}^T$ is generated according to an AR(1) process (2.7) with $\phi_0 = 0$ and $\phi_1 = 0.98$ and where $|\beta_{t1}^0| > 0.5$. The remaining three predictors are allowed to enter and leave the model as time progresses. The regression coefficients $\{\beta_{t2}\}_{t=1}^T$, $\{\beta_{t3}\}_{t=1}^T$ and $\{\beta_{t4}\}_{t=1}^T$ are again generated from an AR(1) process ($\phi_0 = 0$ and $\phi_1 = 0.98$). However, the values are thresholded to zero whenever the absolute value of the process drops below 0.5, creating zero-valued periods. The true sparse series of coefficients are depicted in Figure 4 (black lines).

We begin with the standard approach assuming that all the selection indicators are switched on, i.e. $\gamma_{tj} = 1$ for $t = 0, \ldots, T$ and $j = 1, \ldots, p$. The ASSP construction then boils down to the familiar AR(1) Gaussian state space model. We compute the mode of the joint posterior $\pi(\beta_0, B | Y)$ through standard “coordinate-wise” updates of entire vectors $\beta_t = (\beta_{t1}, \ldots, \beta_{tp})'$. The autoregressive parameters are set to the true values $\phi_0 = 0, \phi_1 = 0.98$ and the conditional Gaussian variance is set to $\lambda_1 = 10(1 - \phi_1^2)$. With the absence of the spike, the estimated series of coefficients cannot achieve sparsity (Figure 4, red lines). Unfortunately, the failure to detect zero directions affects the estimates of the nonzero directions that try to overcompensate for the overfitting. This results in loss of efficiency and poor recovery. The sum of squared differences between the estimate and the
Figure 4: True and recovered time series of regression coefficients from the low-dimensional simulated example with \( p = 6 \)

truth equals \( SE(\hat{B}, B_0) \equiv \sum_{t,j} [\hat{\beta}_{tj} - \beta_{0tj}]^2 = 44.56 \).

With the hope to improve on this recovery, we deploy the ASSP process with a sparsity inducing spike. The hyper-parameters of the slab process are as before (i.e. \( \phi_0 = 0, \phi_1 = 0.98 \) and \( \lambda_1 = 10(1 - \phi_1^2) \)). The remaining tuning parameters are chosen as in Section 4.0.1, i.e. \( \lambda_0 = 1 \) and \( \Theta = 0.9 \). This hyper-parameter choice corresponds to a very mild separation between the stationary spike and slab distributions, and unimodal retrospective and prospective penalties. We deploy the one-step-late EM algorithm outlined in Section 5, initializing the calculation with a zero matrix. We assume that the initial vector \( \beta_{t=0} \) is drawn from the stationary distribution and we estimate it together with all the other parameters, as prescribed in Section 5.

The recovered series have a strikingly different pattern compared to the non-sparse solution (Figure 4, green lines). First, the MAP series is seen to track closely the periods of predictor importance/irrelevance, achieving dynamic variable selection. Second, by harvesting sparsity, the ASSP priors alleviate bias in the nonzero directions, outputting a cleaner representation of the true underlying signal (\( SE(\hat{B}, B_0) = 27.40 \), a marked improvement). The ASSP priors mitigate
overfitting by eliminating noisy coefficients and thereby leaving enough room for the true predictors to capture the trend. As a revealing byproduct, we also obtain the evolving mixing weights determining the relevance of each coefficient at each time. The evolutions of rescaled weights (so that $\theta(0) = 0$) are plotted in Figure 5 (red lines). These companion plots are helpful visualizations of the time-varying sparsity profile.

A similarly adaptive behavior would be also expected from the latent threshold process of Nakajima and West (2013) or the $NGAR$ process of Kalli and Griffin (2014). Both of these methods, however, rely on MCMC. Our solution was obtained with a quick MAP calculation.

6.2 High-dimensional Case

Going further, we consider even more challenging example with $p = 50$ and $T = 100$, yielding a total number of 5000 coefficients to be estimated. The true underlying signal consists of the four series of active regression coefficients introduced earlier in Section 6.1 (Figure 4; black lines in the first four panels). These constitute the only active coefficients in the true regression matrix $B_{50 \times 100}^0$, where the rest of the coefficients are zero. We generate new regressors $x_{ij}$ (iid standard normal) and new
Figure 6: True and recovered time series of regression coefficients from the simulated example with $p = 50$; only first 6 series are plotted

responses from (1.1) with $\sigma_t^2 = 0.25$. At each time point, there is only one univariate observation $y_t$ for every 50 coefficients in $\beta_t$. The estimation challenge will unlikely be met without properly exploiting the underlying sparsity and smoothness.

The recovery obtained with the standard $AR(1)$ Gaussian state space prior (with $\phi_0 = 0$ and $\phi_1 = 0.98$) is far from satisfactory. Plots of the estimated trajectories of the first 6 series (including the 4 active ones) are in Figure 6. By failing to segregate the coefficients as active or inactive, the $AR(1)$ process confuses the source of the signal by smearing it across the redundant covariates. The comparison with Figure 4 is rather striking. With only two irrelevant predictors, the active coefficients in Figure 4 are still recovered reasonably well. Figure 6 illustrates how dramatically the recovery deteriorates when more irrelevant covariates are added. The estimated coefficient evolutions become inconclusive for assessing variable importance, where the coefficient estimates for the relevant variables have been reduced by the increased estimates for the irrelevant variables. This curse of overfitting is exacerbated when $p$ gets even larger.

The ASSP priors serve to overcome this limitation. With a similar prior specification as in
Figure 7: Plots of all the estimated regression coefficients from the higher-dimensional simulated example

Section 6.1 ($\phi_0 = 0$, $\phi_1 = 0.98$, $\Theta = 0.9$ and $\lambda_0 = 0.7$, $\lambda_1 = 10 \times (1 - \phi_1^2)$) we obtain coefficient estimates that are seen to track the true signal far more closely (Figure 6; green lines). Although the recovery is not as accurate as in Figure 4, ASSP priors are still able to tease out the signal even in this very challenging scenario, identifying the active covariates and their active predictive periods. Figure 6 plots only the first 6 coefficient evolutions. Plots of the entire estimated matrix $\hat{B}$ with all 5000 coefficients are in Figure 7. On the left, we have the estimates obtained with the AR(1) process. We can see a marked over-shrinkage of large effects and under-shrinkage of the zero effects. In the middle, we have the ASSP-estimated coefficients with $\lambda_0 = 0.7$ (the same settings as in Figure 6), showing far less bias. With $\lambda_0 = 0.7$, we obtained only 7 false negatives at the expense of 1334 false positives. Fewer false positives can be obtained with larger values $\lambda_0$ (Figure 7(c)), at the expense of more false negatives.

The companion dynamic sparsity plots for the first six series of regression coefficients are compared with the lower-dimensional case in Figure 5 (black lines). Although the two plotted series were obtained on different datasets under different tuning parameters, we can see a level of agreement about the predictive breaks (zero periods).

7 Discussion

This paper introduces a new class of dynamic shrinkage priors. A unique feature of these Autoregressive Spike-and-Slab Process (ASSP) is that their stationary distribution is fully known, where the marginals are characterized by benchmark spike-and-slab priors. A key to obtaining this stabilizing
property is the careful hierarchical construction of adaptive mixing weights that allows them to depend on the lagged value of the process, thereby reflecting sparsity of past coefficients. We propose various versions of ASSP, using Laplace/Gaussian spike/slab distributions. For implementation, we resort to optimization as a faster alternative to posterior sampling. We implement a one-step late algorithm for MAP estimation which iterates over one-site closed-form thresholding rules. Through simulations, we demonstrate that ASSP are well suited for the dual purpose of dynamic variable selection (through thresholding to exact zero) and smoothing (through an autoregressive slab process).

Many variants and improvements are possible on the ASSP prototype constructions. While these go beyond the scope of the present paper, we would like to make the reader aware of generalizations that might be important for future deployment of these priors. The first will be linking the time series priors across the different coordinates. This can be achieved by endowing the importance weight $\Theta$ with a prior distribution, thereby allowing to adapt to the overall sparsity pattern. This elaboration can be embedded within our framework by adding an extra simple step in the algorithm. A modification of Theorem 1 immediately follows, where the marginals are underpinned by the beta-binomial prior as opposed to the Bernoulli prior. The second very important extension will be treating the residual variances $\sigma_t^2$ as random and possibly time-varying. In this paper, we focused primarily on the priors on the regression coefficients. However, the ASSP priors can be deployed in tandem with a stochastic process prior on $\{\sigma_t^2\}_{t=1}^T$ (as e.g. in Kalli and Griffin (2014)). Third, the autoregressive parameters ($\phi_0, \phi_1$) have been assumed fixed and shared across coordinates. To allow for the possibility of treating these as unknown with a prior will be an important extension.

Acknowledgments

This work emerged from my interactions with the Optimization Bayesian Analysis & Decisions (O,BAD) group of the SAMSI Optimization program 2016/17. I would like to thank Mike West for providing useful comments on my presentation, which served as a springboard for this paper. I would also like to thank Ed George for reading through the paper and for providing generous suggestions and comments. This work was supported by the James S. Kemper Foundation Faculty Research Fund at the University of Chicago Booth School of Business.
References


