Dynamic sparsity on time-varying Cholesky-based covariance matrices

Paloma Vaissman Uribe

Tese apresentada
ao
Instituto de Matemática e Estatística
da
Universidade de São Paulo
para
obtenção do título
de
Doutor em Ciências

Programa: Estatística
Orientador: Prof. Dr. Pedro Alberto Morettin
Coorientador: Prof. Dr. Hedibert Freitas Lopes

Durante o desenvolvimento deste trabalho o autor recebeu auxílio financeiro da CAPES/CNPq

São Paulo, julho de 2017

Dynamic sparsity on time-varying Cholesky-based covariance matrices

Esta é a versão original da dissertação/tese elaborada pelo
candidato Paloma Vaissman Uribe, tal como
submetida à Comissão Julgadora.

Acknowledgements

First of all I would like to thank my PhD supervisors, Pedro A. Morettin and Hedibert F. Lopes for their excellent guidance and academic support. In particular, I would like to thank Hedibert for giving me also emotional support during these years and the opportunity to work as a teaching assistant in several disciplines in the area of Econometrics and Time Series within Insper.

Second, I want to thank all my family, specially: my mother for her financial support during the first years of the PhD and for the unconditional love; To my sister, for the moments of leisure, conversations and for having endured all the ends while I had to be absent in difficult moments; To my beloved grandfather, for making me "think big"and for the enormous affection.

I also thank to my colleagues from IME-USP, especially Gilberto, Daniel, Rafael, Cris, Miguel, Julio, Ivan, Isabela, Guaraci, Helder, Lucas, among others: together, we managed to go through so many challenges and good moments. I will always remember the long afternoons of study in the "glass room".

In addition, I am also grateful to all my friends who indirectly contributed to my work, in particular, Marina, Éverton, Flávio, Paula and Denisard; to all the teachers and staff from IME with whom I have taken classes and/ or academic support; to Bruno and co-workers for their patience and encouragement.

Finally, I would like to thank CAPES and CNPq for their financial support.

Abstract

URIBE, P. V. **Dynamic sparsity on time-varying Cholesky-based covariance matrices**. 2017. Tese (Doutorado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2017.

In the present work, we consider variable selection and shrinkage for Gaussian *Dynamic Linear Models* (DLM) within a Bayesian framework. In particular, we propose a novel method that accommodates time-varying sparsity, based on an extension of spike-and-slab priors for dynamic models. This is done by assigning appropriate priors for the time-varying coefficients' variances, extending the previous work of Ishwaran and Rao (2005). Our approach is similar to the *Normal Gamma Autoregressive* (NGAR) process of Kalli and Griffin (2014), nevertheless, we assume a Markov switching structure for the process variances instead of a *Gamma Autoregressive* (GAR) process. Furthermore, we investigate different priors, including the common Inverted gamma prior for the process variances, and other mixture prior distributions such as Gamma priors for both the spike and the slab, which leads to a mixture of Normal-Gammas priors (Griffin et al. (2010)) for the coefficients and also different distributions for the spike and the slab. In this sense, our prior can be view as a dynamic variable selection prior which induces either smoothness (through the slab) or shrinkage towards zero (through the spike) at each time point. The MCMC method used for posterior computation uses Markov latent variables that can assume binary regimes at each time point to generate the coefficients' variances. In that way, our model is a dynamic mixture model, thus, we could use the algorithm of Gerlach et al. (2000) to generate the latent processes without conditioning on the states. Finally, our approach is exemplified through simulated examples and a real data application.

Keywords: Cholesky decomposition, dynamic models, Normal-Gamma prior, spike-and-slab priors, high-dimensional data, scale mixture of Normals.

Resumo

URIBE, P. V. **Esparsidade dinâmica em matrizes de covariância variantes no tempo via decomposição de Cholesky**. 2017. Tese (Doutorado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2017.

No presente trabalho são apresentados diversos métodos de seleção de variáveis e encolhimento para modelos lineares dinâmicos Gaussianos sob a perspectiva Bayesiana. Em particular, propomos um novo método o qual induz esparsidade dinâmica em modelos de regressão linear com coeficientes variantes no tempo. Isso é feito através da especificação de prioris *spike-and-slab* para as variâncias dos coeficientes de variação do tempo, estendendo o trabalho anterior de Ishwaran and Rao (2005). A abordagem é semelhante ao processo definido em Kalli and Griffin (2014), no entanto, assumimos uma estrutura *Markov switching* para as variâncias ao invés de um processo Gama autoregressivo. Além disso, investigamos diferentes priores, incluindo uma mistura de distribuições Gama Inversa, bastante utilizada para variâncias, além de outras misturas de distribuições, como a Gama, que gera a priori conhecida como Normal-Gama para os coeficientes (Griffin et al. (2010)). Nesse sentido, o modelo proposto pode ser visto como uma seleção de variável dinâmica em que os coeficientes podem assumir valores diferentes de zero seguindo uma distribuição mais dispersa (através do *slab*) ou encolhimento em direção a zero (através do *spike*) em cada ponto do tempo. O esquema MCMC usado para simular a posteriori utiliza variáveis latentes Markovianas que podem assumir regimes binários em cada ponto de tempo para gerar as variâncias dos coeficientes. Dessa forma, o modelo é um modelo de mistura dinâmica, portanto, para gerar as variáveis latentes, utilizamos o algoritmo de Gerlach et al. (2000), que permite gerar essas variáveis sem condicionamento nos estados (coeficientes variantes no tempo). A abordagem é exemplificada através de exemplos simulados e uma aplicação de dados reais.

Palavras-chave: Decomposição de Cholesky, modelos dinâmicos, priori Normal-Gama, priori spike-and-slab, misturas de escala Normal.

Contents

List of Symbols

$\mathcal{B}(\alpha, \beta$	Beta distribution with parameters α and β
$Bin(n, p)$	Binomial distribution with sample size n and probability p
$\mathcal{E}(\alpha)$	Exponential distribution with mean $1/\alpha$
$\delta_{\upsilon}(.)$	discrete measure concentrated at value υ
$\mathcal{G}(\alpha, \beta)$	Gamma distribution with shape parameter a and rate parameter b
$\mathcal{IG}(\alpha, \beta)$	Inverse Gamma distribution with shape parameter a and scale parameter b
$\mathcal{N}(\mu, \sigma^2)$	Normal distribution with mean μ and variance σ^2
$t_{\nu}(\mu, \sigma^2)$	Student-t distribution with location parameter μ, scale σ^2 and ν degrees of freedom
$Lap(x)$	Laplace distribution with mean 0 and scale parameter x

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

Over the past few decades, advances in computational processing have encouraged the proliferation of massive datasets, bringing new perspectives and challenges to statistical research due to high-dimensionality issue. In this sense, regularization and variable selection techniques have become even more relevant to induce sparsity and solve ill-posed problems. A few years ago, Hastie et al. (2001) coined the informal *Bet on Sparsity* principle, which encourages the use of procedures that do well in sparse problems for high-dimensional problems, since no procedure does well in dense problems. Actually, they have shown that for a dense problem, where all the numerous coefficient where different from zero, and/or there is a high *Noise-to-Signal Ratio* (NSR), both the former *ridge regression* procedure of Hoerl and Kennard (1970) and the *least absolute shrinkage and selection operator* (lasso) from Tibshirani (1996) do poorly in terms of prediction. In contrast, famous statistician Andrew Gelman has questioned this principle in his blog [1] saying that in fact the world is non-sparse and the relations between variables are very complex. Therefore in a dense setting as we encounter in Social Science research one is not actually interested in "recovering the underlying model."

Despite the different point of views, sparsity makes sense and appear frequently in many situations. For instance, in several areas of research, such as genetics, finance and neuroscience, high-dimensional data are collected daily and there is great interest in establishing measures of association or dependence between variables. For example, the estimation of correlation or measures of dependency among several regions of interest (ROI) of the brain is of great importance in studies of brain connectivity involving functional magnetic resonance imaging data. In finance, the solution to the portfolio optimization problem between risk and expected returns needs the inverse of the covariance matrix among the assets, which may be numerous given the preference for portfolio diversification.

In both mentioned situations, a sparse setting is reliable and even desired for practical reasons. The fact is that estimating a covariance matrix among many variables, considering relatively few observations of each variable is a difficult problem. Let $\boldsymbol{X}$ be a data matrix $(n \times p)$, where n is the number of sample observations and p the number of variables. We know that to estimate a complete or unconditional covariance matrix, we need to estimate $p(p-1)/2$ parameters, which may be really complex if p is big. Another difficulty that occurs in high-dimensional problems is that many methods, such as least squares, require the inverse

[1] http://andrewgelman.com/2013/12/16/whither-the-bet-on-sparsity-principle-in-a-nonsparse-world/

calculation of $\boldsymbol{X}'\boldsymbol{X}$, which is obviously singular when $p \gg n$. If we think about time-varying covariance matrices, the problem gets even harder as now we have to estimate $p(p-1)n/2$ parameters.

In science, Occam's razor principle is used as a heuristic guide in the development of theoretical models. Although it is not considered an irrefutable principle of logic or a scientific result, the preference for simplicity in the scientific method is based on the falsifiability criterion as long as simpler theories are preferable to more complex ones because they are more testable. Applied to statistical analysis, this implies that the less dense model that fits the data is best as unnecessary predictors will add noise to the estimation and, besides that, degrees of freedom will be wasted. For these reasons, regularization and variable selection techniques are particularly encouraged as they involve some form of dimensionality reduction, making regression problems less complex.

The problem of variable selection refers to the statistical endeavor of selecting a subset of observed characteristics, which collectively provide a good description of an observed phenomenon. Of particular interest are settings where such a subset is parsimonious. In this sense, variable selection acts as a form of model selection within a regression framework, where models differ in their configuration of the contributing variables. In frequentist approach, stepwise regression and criterion based methods are very well known techniques for model selection. In Bayesian framework, variable selection is commonly based on spike-and-slab priors for regression coefficients, being the focus of this thesis. Seminal papers on this topic are Mitchell and Beauchamp (1988) and George and McCulloch (1993), where the latter is called the *stochastic search variable selection* (SSVS) method.

While variable selection stands for model selection, regularization methods uses the entire dictionary of variables but restrict the coefficients. In the frequentist framework, this is the penalized maximum likelihood estimation, such as the lasso Tibshirani (1996) and ridge regression of *ridge regression*, where the former imposes a ℓ_2 penalty on the regression coefficients, while the latter works with the ℓ_1 penalty. As noted in Hastie et al. (2001), ridge regression is a simple example of a regularization approach, while the lasso is both a regularization and selection method, because it can be proven that lasso actually induces sparsity even though it has some bias problems, as it also over shrinks non-zero coefficients in some cases. In the Bayesian context, regularization within regression problems is equivalent to assigning shrinking priors to the coefficients.

1.1 Outline

In this thesis, we focus on Bayesian methods, discussing several shrinking priors for regularization purposes and well as spike-and-slab priors for variable selection. The aim of the work is to derive a Bayesian variable selection method within the framework of the Gaussian *Dynamic Linear Models* (DLM). Furthermore, besides estimating and shrinking the time-varying coefficients horizontally, we want to derive a method that accommodates time-varying sparsity, that is, where the subset of relevant predictors also change over time (vertical sparsity).

First, it is worth explain in details the concepts of vertical and horizontal sparsity. To make these concepts even more clear, Table 1.1 above illustrates a time-varying sparsity pattern for

$q = 5$ potential predictors and $t = 1, .., 12$ (months). One can see that the regressor $\boldsymbol{X}_1$ always have an irrelevant effect such as it can be estimated as being zero at all times (horizontal sparsity). Nevertheless, variables $\boldsymbol{X}_2$,$\boldsymbol{X}_3$ and $\boldsymbol{X}_4$ are relevant in some quarters and irrelevant in others, therefore the subset of non negligible variables varies over months (vertical sparsity).

	jan	feb	mar	apr	may	jun	jul	aug	sep	oct	nov	dec
x_0	$\beta_{1,1}$	$\beta_{1,2}$	$\beta_{1,3}$	$\beta_{1,4}$	$\beta_{1,5}$	$\beta_{1,6}$	$\beta_{1,7}$	$\beta_{1,8}$	$\beta_{1,9}$	$\beta_{1,10}$	$\beta_{1,11}$	$\beta_{1,12}$
x_1	0	0	0	0	0	0	0	0	0	0	0	0
x_2	$\beta_{3,1}$	$\beta_{3,2}$	$\beta_{3,3}$	$\beta_{3,4}$	$\beta_{3,5}$	0	0	0	$\beta_{3,9}$	$\beta_{3,10}$	$\beta_{3,11}$	$\beta_{3,12}$
x_3	0	0	$\beta_{4,3}$	$\beta_{4,4}$	$\beta_{4,5}$	$\beta_{4,6}$	$\beta_{4,7}$	$\beta_{4,8}$	$\beta_{4,9}$	$\beta_{4,10}$	$\beta_{4,11}$	0
x_4	$\beta_{5,1}$	$\beta_{5,2}$	$\beta_{5,3}$	$\beta_{5,4}$	$\beta_{5,5}$	0	0	0	$\beta_{5,9}$	$\beta_{5,10}$	$\beta_{5,11}$	$\beta_{5,12}$

Table 1.1: *Horizontal and vertical sparsity example*

Our method is based on placing spike-and-slab priors on the time-varying variances' of the time-varying coefficients, extending the previous work of Ishwaran and Rao (2005). That is, for each time point t, each $\beta_{j,t}$ from the linear regression with $j = 1, ..., q$ predictors is (marginally) normally distributed as

$$\beta_{j,t}|\psi_{j,t} \sim \mathcal{N}(0, \psi_{j,t}),$$

for $t = 1, ..., T$ where each parameter $\psi_{j,t} = K_{j,t}\tau_j^2$, conditional on the previous value $K_{j,t-1}$ is a finite mixture

$$\psi_{j,t}|K_{j,t-1} \propto \omega p_{slab}(\psi_{j,t}) + (1-\omega)p_{spike}(\psi_{j,t}),$$

where the weight $\omega = f(K_{j,t-1})$ is a function of the previous value of $K_{j,t-1}$ and the latent variables $K_{j,1}, .., K_{j,T}$ evolves as a Markov switching process. The basic idea is that each variance $\psi_{j,t}$ is modeled as having come either from a distribution with most (or all) of its mass concentrated around zero ($p_{spike}(\psi_{j,t})$), or from a comparably diffuse distribution with mass spread out over a large range of values ($p_{slab}(\psi_{j,t})$).

We investigate different prior distributions for $p_{spike}(\psi_{j,t})$ and $p_{slab}(\psi_{j,t})$, including the common Inverted Gamma prior for the process variances, and other mixture prior distributions such as Gamma priors for both the spike and the slab, which leads to a mixture of Normal-Gammas priors (Griffin et al. (2010)) for each $\beta_{j,t}$. In this sense, our prior can be view as a dynamic variable selection prior which induces either smoothness (through the slab) or shrinkage towards zero (through the spike) at each time point.

The posterior computation is done by a Gibbs Sampler with a Metropolis step. The Markov latent variables $K_{j,1}, .., K_{j,T}$ can assume binary regimes at each time point and they generate the coefficients' variances $\psi_{j,1}, ..., \psi_{j,T}$, which in turn shrink the coefficients. In that way, our model is a dynamic mixture model, thus, we could use the algorithm of Gerlach et al. (2000) to generate the latent processes without conditioning on the states. Furthermore, the states $\beta_{j,t}$ are generated using the *Forward Filtering Backward Sampling* (FFBS) algorithm (Carter and Kohn (1994); Frühwirth-Schnatter (1994)). The other parameters from the regression are also sampled within the MCMC scheme, in a full Bayes strategy.

1.2 Organization

The rest of this work is organized as follows:

- Chapter 2 is devoted to study of the Gaussian linear model and the regularization and variable selection methods applied to linear regression models. First, the Gaussian linear model is presented and several estimation methods are discussed such as the classical ordinary least squares as well as the Bayesian inference of the Gaussian linear model. Second, some regularization methods are introduced in Section 2.2, such as the *lasso* (Tibshirani (1996)), the *Bayesian lasso* (Park and Casella (2008)), the *ridge regression* (Hoerl and Kennard (1970)) and the *elastic net* (Zou and Hastie (2005)). Finally, Bayesian methods of variable selection are presented as an alternative to the classical stepwise regression approaches in Section 2.3. Specifically, we discuss the *stochastic search variable selection* method of George and McCulloch (1993) and the *stochastic variable selection* model of Ishwaran and Rao (2005), both based on spike-and-slab priors. Then, we generalize the spike-and-slab prior by allowing different combinations of distributions, following the work of Frühwirth-Schnatter and Wagner (2011).

- Chapter 3 is dedicated to presenting the general Gaussian Dynamic Linear Model theory (DLM) and the existing literature of regularization and variable selection methods applied to Time Varying Parameter (TVP) models. Specifically, in Section 3.1, we discuss the basic formulation of state space models and the structure of the recursive computations for estimation and prediction. Then, in Section 3.2, the specific case of the Gaussian DLM is presented as well as the the *Kalman filter*, the *Kalman smoother* and the *Forward Filtering Backward Sampling* (FFBS) algorithm for Gaussian DLMs. Finally, we discuss some existing regularization methods for TVP models in Section 3.3 that have influenced our proposed model which will be discussed in Chapter 5.

- Chapter 4 is allocated to discussing some existing methods for sparse covariance modeling based on the regularization of the linear regressions which result from the covariance matrix decompositions. The objective is to briefly discuss the various decompositions of the covariance matrix that make the problem of estimation of a matrix into a linear regression problem, especially in the case of high-dimensional problems. In particular, we choose the modified Cholesky decomposition for the applications, because of its natural interpretation and practical appeal. Thus, in Section 4.1 we present the modified Cholesky and other covariance matrix decompositions. Then, Section 4.2 is dedicated to presenting existing frequentist methods for regularizing the Cholesky factor such as the approach of Huang et al. (2006), based on the lasso regularization, the *Adaptative Banding with a nested lasso penalty* (AB) of Levina et al. (2008) and the *Forward Adaptative Banding* of Leng and Li (2011). Lastly, we present some simulated and real data examples and compare with the Bayesian regularization of the Cholesky linear regressions based on the Normal-Gamma prior.

- Finally, in Chapter 5 we propose a new method that accommodates time-varying sparsity, based on spike-and-slab priors. In Section 5.1, we present the formulation of the proposed model and its posterior inference. In Section 5.2, some simulated examples, including one where we simulate and obtain a time-varying covariance matrix using the Cholesky decomposition, are given. An empirical example using inflation data is given in Section 5.3.

Chapter 2

Regularization and variable selection in the Gaussian linear model

In this chapter we consider some approaches for extending the Gaussian linear model framework. In particular, we discuss several regularization and variable selection methods applied to linear regression models.

First, the Gaussian linear model is presented and several estimation methods are discussed in Section 2.1, such as the classical ordinary least squares and the maximum likelihood approaches, as well as the Bayesian inference of the Gaussian linear model.

Second, some regularization methods for linear regression models are introduced in Section 2.2, such as the *lasso* (Tibshirani (1996)), the *Bayesian lasso* (Park and Casella (2008)), the *ridge regression* (Hoerl and Kennard (1970)) and the *elastic net* (Zou and Hastie (2005)). All these methods are discussed both from the classical point of view, which is based on the idea of penalized maximization (or minimization), and from the Bayesian perspective, which is based on assigning priors for the regression coefficients, mostly within the class of scaled mixture of normals (see, e.g.,West (1987)).

Finally, Bayesian methods of variable selection are presented as an alternative to the classical stepwise regression approaches in Section 2.3. Specifically, we discuss the *stochastic search variable selection* method of George and McCulloch (1993) and the *stochastic variable selection* model of Ishwaran and Rao (2005), both based on spike-and-slab priors. Then, we generalize the spike-and-slab prior by allowing different combinations of distributions, following the work of Frühwirth-Schnatter and Wagner (2011).

2.1 The Gaussian linear model

The Gaussian linear model is defined by

$$\boldsymbol{y} = \beta_0 \mathbf{1} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}), \tag{2.1}$$

where $\boldsymbol{y}$ denotes a n-dimensional vector of continuous responses, β_0 is the intercept, $\boldsymbol{\beta}$ is a q-dimensional vector of regression coefficients associated with covariates, $\boldsymbol{X}$ is a $(n \times q)$ design matrix with each column representing a covariate, and $\boldsymbol{\varepsilon}$ is multivariate Gaussian (i.e., $\mathcal{N}$ denotes the Gaussian distribution). Note that each component ε_i of $\boldsymbol{\varepsilon}$, for $i = 1, ..., n$, is an independent and identically distributed Gaussian error with zero mean and variance σ^2.

Notice that, generally and without loss of generalization, the design matrix $\boldsymbol{X}$ is often standardized, namely, the columns of $\boldsymbol{X}$ are centered and scaled (for each column $j = 1, ..., q$, we have $\mathbb{E}(\boldsymbol{X}_j) = \boldsymbol{0}$ and $\text{Var}(\boldsymbol{X}_j) = 1$) so that the estimated values of $\boldsymbol{\beta}$ are truly comparable. In addition, because we can define $\alpha = \bar{y}$, where $\bar{y} = \sum_{i=1}^{n} y_i/n$, if we replace the responses y_i by their centered values $y_i - \bar{y}$, for $i = 1, ..., n$, then we can simply ignore the intercept β_0.

From Equation (2.1), rearranging the design matrix $\boldsymbol{X}$ and the coefficients' vector $\boldsymbol{\beta}$, we may rewrite the model as

$$(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}, \sigma^2) \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}), \tag{2.2}$$

where now $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_q)$ and $\boldsymbol{X}$ is a $(n \times (q+1))$ matrix (we add a column of 1s into $\boldsymbol{X}$).

2.1.1 The classical approach

Consider the Gaussian linear model presented in Equation (2.2). A model fitting procedure produces the vector of estimates $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_q)$. The ordinary least squares (OLS) estimates are obtained by minimizing the residual sum of squares

$$RSS(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2,$$

that is,

$$\hat{\boldsymbol{\beta}}_{OLS} = \arg\min_{\boldsymbol{\beta}}\{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2\},$$

which results in the unique solution

$$\hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}. \tag{2.3}$$

Note that under Gaussian errors, the OLS solution is equivalent to the classical maximum likelihood estimate (MLE), which is obtained by maximizing the likelihood

$$\mathcal{L}(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right\}.$$

It is well known that the OLS/MLE estimator is unbiased and that the sampling distribution considering the Gaussian linear model is

$$\hat{\boldsymbol{\beta}}_{OLS} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}),$$

while assuming that σ^2 is known. Nevertheless, in most cases, σ^2 in unknown and has to be estimated using the mean squared residual error

$$s^2 = \frac{n}{n-q}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2,$$

in which case $\hat{\boldsymbol{\beta}}_{OLS}$ has a multivariate t sampling distribution centered on $\boldsymbol{\beta}$.

2.1.2 Bayesian inference in the Gaussian linear model

We describe two cases of Bayesian inference with conjugate priors for the regression model: inference on the regression coefficients $\boldsymbol{\beta}$, assuming that σ^2 is known, and inference on $\boldsymbol{\beta}$ and σ^2, assuming that both are unknown.

Inference on $\boldsymbol{\beta}$, σ^2 known. A typical approach is to introduce a conjugate Gaussian prior for the coefficients, for instance

$$\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{S}_0), \tag{2.4}$$

while assuming that σ^2 is known. The posterior is then

$$(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X}, \sigma^2) \propto \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) - \frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\mu}_0)'\boldsymbol{S}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu}_0)\right\},$$

where, completing the square inside the exponent, we get the following posterior

$$(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X}, \sigma^2) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{S}), \tag{2.5}$$

with

$$\begin{aligned} \boldsymbol{S} &= \sigma^2(\boldsymbol{X}'\boldsymbol{X} + \sigma^2\boldsymbol{S}_0^{-1})^{-1}, \\ \boldsymbol{\mu} &= \boldsymbol{S}(\boldsymbol{S}_0^{-1}\boldsymbol{\mu}_0 + \frac{1}{\sigma^2}\boldsymbol{X}'y). \end{aligned}$$

Note that when $\boldsymbol{S}_0^{-1} \to \boldsymbol{0}$, $\boldsymbol{\mu} \to \hat{\boldsymbol{\beta}}_{OLS}$. The same happens when we place a non-informative (the so-called Jeffrey's) prior on $\boldsymbol{\beta}$. That is, if $\boldsymbol{\beta} \propto \boldsymbol{1}$, then $(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X}, \sigma^2) \sim \mathcal{N}(\hat{\boldsymbol{\beta}}_{OLS}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})$.

Inference on both $\boldsymbol{\beta}$ and σ^2. As we rarely know σ^2, another typical approach is placing the hierarchical prior

$$\boldsymbol{\beta}|\sigma^2 \sim \mathcal{N}(\boldsymbol{\mu}_0, \sigma^2\boldsymbol{S}_0), \qquad \sigma^2 \sim \mathcal{IG}(a, b), \tag{2.6}$$

where $\mathcal{IG}$ denotes the Inverse-Gamma distribution with shape a and scale b.

Note that the hierarchical structure in (2.6) is the same of placing a conjugate prior for both $(\boldsymbol{\beta}, \sigma^2)$, i.e.,

$$p(\boldsymbol{\beta}, \sigma^2) = p(\boldsymbol{\beta}|\sigma^2)p(\sigma^2) = \mathcal{N}(\boldsymbol{\mu}_0, \sigma^2\boldsymbol{S}_0)IG(a, b),$$

where this structure is called the Normal Inverse-Gamma (NIG) prior with parameters $(\boldsymbol{\mu}_0, \boldsymbol{S}_0, a, b)$. The resulting posterior distribution is

$$(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}, \boldsymbol{X}) \sim \mathcal{N}(\boldsymbol{\mu}_n, \sigma^2\boldsymbol{S}_n^{-1})\mathcal{IG}(a_n, b_n), \tag{2.7}$$

where

$$\begin{aligned}
a_n &= a + n/2,\\
b_n &= b + \frac{1}{2}(\boldsymbol{y}'\boldsymbol{y} + \boldsymbol{\mu}_0'\boldsymbol{S}_0^{-1}\boldsymbol{\mu}_0 - \boldsymbol{\mu}_n'\boldsymbol{S}_n\boldsymbol{\mu}_n),\\
\boldsymbol{\mu}_n &= (\boldsymbol{S}_0^{-1} + \boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{S}_0^{-1}\boldsymbol{\mu}_0 + \boldsymbol{X}'\boldsymbol{y}),\\
\boldsymbol{S}_n &= \boldsymbol{S}_0^{-1} + \boldsymbol{X}'\boldsymbol{X}.
\end{aligned}$$

An interesting analytic form results from integrating out σ^2 from the joint posterior density in (2.7). In this case, the marginal posterior of $\boldsymbol{\beta}$ follows a multivariate t-distribution

$$(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X}) = \int p(\boldsymbol{\beta},\sigma^2|\boldsymbol{y},\boldsymbol{X})d\sigma^2 \sim t_\nu(\boldsymbol{\mu}^*,\boldsymbol{S}^*),$$

with

$$\begin{aligned}
\boldsymbol{\mu}^* &= \boldsymbol{\mu}_n,\\
\boldsymbol{S}^* &= \frac{b_n}{a_n}\boldsymbol{S}_n,
\end{aligned}$$

and $\nu = 2a_n$ degrees of freedom.

g-prior & ridge prior. In practice, NIG prior is too informative and has too many constants. People often prefer to use the g-prior introduced by Zellner (1986)

$$\boldsymbol{\beta}|\sigma^2 \sim \mathcal{N}(\boldsymbol{0}, g\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}), \tag{2.8}$$

where for choosing g one could use information criteria BIC, empirical Bayes or full Bayes strategies.

Note that (2.8) is equivalent to the NIG prior presented in (2.6) with $\boldsymbol{\mu}_0 = \boldsymbol{0}$, $\boldsymbol{S}_0 = g(\boldsymbol{X}'\boldsymbol{X})^{-1}$ and $a \to 0, b \to 0$, i.e., the Jeffrey's prior $p(\sigma^2) \propto \frac{1}{\sigma^2}$. The posterior density of $\boldsymbol{\beta}$ is Gaussian

$$(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X},\sigma^2) \sim \mathcal{N}\left(\frac{g}{g+1}\hat{\boldsymbol{\beta}}_{OLS}, \frac{g}{g+1}\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\right),$$

where letting $g \to \infty$, we recover the frequentist $\hat{\boldsymbol{\beta}}_{OLS}$ from Equation (2.3).

Finally, another famous Bayesian approach for linear Gaussian model is the ridge prior, which is taken to be

$$\boldsymbol{\beta} \sim \mathcal{N}(0, \sigma^2\lambda^{-1}\boldsymbol{I}), \tag{2.9}$$

where $\lambda > 0$ and $\sigma^2 > 0$ are known scalars. Note that this is equivalent to (2.4), where $\boldsymbol{S}_0 = \sigma^2\lambda^{-1}\boldsymbol{I}$ and σ^2 is known. Thus, the posterior is Gaussian with posterior mean $\boldsymbol{\mu}$ given by

$$\boldsymbol{\mu} = (\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y}.$$

In the next section, we will give more details about ridge prior and its frequentist analogue.

2.2 Regularization

It is well known that OLS often does poorly on both prediction accuracy and interpretation, as noted in Tibshirani (1996) and in Zou and Hastie (2005). Interpretation stands for the preference for parsimony, in the sense that simpler models put more light on the relationship between the response and covariates. Prediction accuracy is related to the bias-variance trade-off.

Consider the Gaussian linear model defined in the Equation (2.2). For simplicity, assume throughout this section that the covariates from matrix $\boldsymbol{X}$ are standardized and the response $\boldsymbol{y}$ is centered, so that we can ignore the intercept. Let $\hat{\boldsymbol{\beta}}$ be an estimate for $\boldsymbol{\beta}$ and $\boldsymbol{X} = \boldsymbol{X}_0$ be a known particular target point for prediction. Suppose we have an estimator for $\boldsymbol{y}$ equal to $\hat{f}(\boldsymbol{X}) = \boldsymbol{X}\hat{\boldsymbol{\beta}}$. Then, the prediction error $pe(\boldsymbol{X}_0)$ is

$$\begin{aligned} pe(\boldsymbol{X}_0) &= \mathbb{E}_{y|X=X_0}\left\{\left(\boldsymbol{y} - \hat{f}(\boldsymbol{X})\right)^2 |\boldsymbol{X} = \boldsymbol{X}_0\right\} \\ &= \sigma^2 + \left(\boldsymbol{y} - \hat{f}(\boldsymbol{X}_0)\right)^2 + \text{Var}\left(\hat{f}(\boldsymbol{X}_0)\right). \end{aligned}$$

Such a decomposition is known as the bias-variance trade-off. Although the OLS/ MLE estimator has the smallest variance among all unbiased estimators accordingly to the Gauss Markov Theorem, an estimator with slight bias but smaller variance could be preferable, leading to a substantial decrease in prediction error. As the model complexity rises (more terms included), OLS estimates suffer from higher variance.

Modern statistics allows for this trade-off between bias and variance through regularization methods, which encourages simpler models because the space of values of $\hat{\boldsymbol{\beta}}$ considered is smaller. This is consistent with the sparsity principle which assumes that only a small number of predictors contribute to the response. Intuitively, regularization prevents overfitting, leading to better generalization.

Another important issue concerning the OLS/MLE estimate is that it is undefined in high-dimensional problems, where the number of variables is much greater than the number of observations ($q \gg n$). In this case, $\boldsymbol{X}'\boldsymbol{X}$ is singular or not well-conditioned.

In general terms, the notion of regularization summarizes approaches that allow to solve ill-posed problems, such as those which arises from high-dimensional data, or to prevent overfitting. Hence, the purpose of regularization is introducing additional information that allow to characterize useful solutions for $\boldsymbol{\beta}$, inducing models to be sparse or introducing a group structure into the problem.

2.2.1 The ridge regression and the lasso: classical and Bayesian approaches

One of the first classical regularization approach was the *ridge regression* from Hoerl and Kennard (1970) or Tikhonov regularization, which solution is given by

$$\hat{\boldsymbol{\beta}}_{ridge} = \arg\min_{\beta}\left\{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda\,\|\boldsymbol{\beta}\|_2^2\right\}, \tag{2.10}$$

where $\lambda \geq 0$ is a regularization parameter controlling the length of the vector of regression coefficients. This is equivalent to saying that the ridge estimate for $\boldsymbol{\beta}$ is the optimal solution obtained by minimizing $\sum_{i=1}^{n}(y_i - \boldsymbol{X}_i\boldsymbol{\beta})^2$ subject to $\sum_{j=1}^{q}\boldsymbol{\beta}_j^2 \leq t$, where $t \geq 0$ is a tuning parameter. That is, the ridge estimate is a penalized least squares method imposing a ℓ_2 penalty on the regression coefficients. The unique ridge solution is

$$\hat{\boldsymbol{\beta}}_{ridge} = (\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y}. \tag{2.11}$$

Note that for each λ we have a solution. If $\lambda \to 0$, we obtain the least squares solution from Equation (2.3) and if $\lambda \to \infty$, we have $\hat{\boldsymbol{\beta}}_{ridge} = 0$. In the original paper, a procedure called ridge traces was discussed, i.e., in order to choose λ, one should plot the estimates against λ and choose the one for which the coefficients are not rapidly changing and have sensible signs. This procedure was heavily criticized, so now the standard practice is to use cross validation for choosing λ. It can be proven that

$$\mathbb{E}(\hat{\boldsymbol{\beta}}_{ridge}) = \boldsymbol{Z}^{-1}\boldsymbol{\beta}$$
$$\text{Var}(\hat{\boldsymbol{\beta}}_{ridge}) = \sigma^2 \boldsymbol{Z}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{Z}',$$

where $\boldsymbol{Z} = (I + \lambda(\boldsymbol{X}'\boldsymbol{X})^{-1})^{-1}$, which means that the ridge estimator is biased. The total variance decreases as λ increases, while the square bias increases with λ. This illustrates the bias-variance trade-off of the ridge estimate.

Even though the *ridge regression* solves problems when there are more variables than observations ($q > n$), it is worth noting that it doesn't induce sparsity solutions, which forces to zero the smaller coefficients, but keeps the bigger ones around.

In this sense, another classical regularization approach for the Gaussian linear regression problem has arisen to fix this drawback: the *least absolute shrinkage and selection operator* (lasso) from Tibshirani (1996). The lasso estimate is defined by

$$\hat{\beta}_{lasso} = \arg\min_{\beta}\{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|_1\}, \tag{2.12}$$

where $\lambda \geq 0$ is a regularization parameter that determines the impact of penalty function. Again, this is equivalent to saying that the lasso estimate for $\boldsymbol{\beta}$ is the optimal solution obtained by minimizing $\sum_{i=n}^{q}(y_i - \boldsymbol{X}_i\boldsymbol{\beta})^2$ subject to $\sum_{j=1}^{q}|\beta_j| \leq t$, where $t \geq 0$ is a tuning parameter. That is, the lasso is a penalized least squares method imposing a ℓ_1 penalty on the regression coefficients.

Note that if $t = \sum_{j=1}^{q}|\hat{\beta}_j^{OLS}|$, then $\lambda = 0$ and we obtain no shrinkage. Therefore, only when $t < \sum_{j=1}^{q}|\hat{\beta}_j^{OLS}|$ will cause shrinkage of the solutions towards zero, and some of the coefficients may be exactly equal to zero. This explains why lasso is also considered a variable selection method by some.

Unlike the ridge estimate, the lasso estimate has no closed form since it is a non linear and non-differentiable function of the response values even for a fixed value of t. Original implementation involves quadratic programming techniques from convex optimization. Again, the lasso shrinking parameter λ is usually chosen by cross validation methods.

Many other sparsity-inducing penalties such as the *elastic net* from Zou and Hastie (2005) and the *adaptive lasso* from Zou (2006) have similar structures, with penalty functions equal to, respectively

$$\lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \|\boldsymbol{\beta}\|_2^2, \lambda_1 \geq 0, \lambda_2 \geq 0 \tag{2.13}$$

and

$$\lambda \sum_{j=1}^{q} |\beta_j| / |\hat{\beta}_j^{OLS}|, \, \lambda \geq 0. \tag{2.14}$$

While the frequentist approaches are based on penalized optimizations, Bayesian regularization for the linear regression can be formalized through the conditional distribution $p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\phi})$, where $\boldsymbol{\phi}$ is a parameter vector, comprising, for example, the error variance σ^2. The regularization is achieved by specifying appropriate informative priors $p(\boldsymbol{\beta}|\boldsymbol{\theta})$, where the hyperparameter vector $\boldsymbol{\theta}$ includes parameters controlling shrinkage properties. The model is completed by assuming hyperpriors $p(\boldsymbol{\theta})$ and $p(\boldsymbol{\phi})$ and the inference is based on the posterior $p(\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\phi}) p(\boldsymbol{\beta}|\boldsymbol{\theta}) p(\theta) p(\boldsymbol{\phi})$. It can be shown that if $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ are fixed, then the posterior mode or the maximum a posteriori estimate (MAP)

$$\arg\max_{\beta} \{ p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\phi}) p(\boldsymbol{\beta}|\boldsymbol{\theta}) \}$$

is equivalent to penalizing the log-likelihood $\log p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\phi})$ with penalty equal to the (minus) log prior $\log p(\boldsymbol{\beta}|\boldsymbol{\theta})$.

In fact, this was pointed out by Tibshirani (1996), who also noted that $|\beta_j|$ is proportional to the log density of the double exponential distribution. As a result, one can derive the lasso estimate as the Bayes posterior mode under independent Laplace or Double-Exponential priors for the components of $\boldsymbol{\beta}$. That is,

$$\beta_j \overset{\text{ind}}{\sim} \frac{1}{2\tau} \exp\left(-\frac{|\beta_j|}{\tau} \right), \tag{2.15}$$

for $j = 1, ..., q$, with $\tau = 1/\lambda$. Here, λ is the shrinking parameter from the lasso penalty.

In general, practically all shrinking priors are defined hierarchically as a scaled mixture of normals (SMN). For details on SMN distributions, see West (1987). The SMN class has the following general structure

$$\beta_j | \psi_j \sim \mathcal{N}(0, \psi_j), \qquad \psi_j | \theta \sim p(\psi_j | \boldsymbol{\theta}), \tag{2.16}$$

where β_i and β_j are independent for any $i, j \in \{1, ..., q\}$ and ψ_j depends on the vector of hyperparameters $\boldsymbol{\theta}$. Note that the marginal distribution of β_j

$$p(\beta_j | \boldsymbol{\theta}) = \int p(\beta_j | \psi_j) p(\psi_j | \boldsymbol{\theta}) d\psi_j$$

is non Gaussian, and can assume many forms depending on the mixing distribution $p(\psi_j | \boldsymbol{\theta})$.

A famous form arises when the mixing distribution is Exponential. As pointed out in Park and Casella (2008), the Laplace prior from (2.15) can be represented by as a scale mixture of normals with the following hierarchic specification

$$(\beta_j | \psi_j) \overset{\text{ind}}{\sim} \mathcal{N}(0, \psi_j), \qquad \psi_j | \lambda \sim \mathcal{E}\left(\frac{\lambda^2}{2} \right), \tag{2.17}$$

for $j = 1, ..., q$, where $\mathcal{E}(\alpha)$ denotes the Exponential distribution with mean $1/\alpha$. Thus, marginally, β_j follows a Laplace distribution with parameter λ, that is, $p(\beta_j) \propto \exp(-\lambda |\beta_j|)$.

In fact, Park and Casella (2008) considered a conditional (on σ^2) Laplace prior specification to guarantee a unimodal full posterior, that is

$$\beta_j|\sigma^2 \overset{\text{ind}}{\sim} \frac{\lambda}{2\sqrt{\sigma^2}} e^{-\lambda|\beta_j|/\sqrt{\sigma^2}} \tag{2.18}$$

for $j = 1, ..., q$, where a non informative scale-invariant marginal prior is assigned to σ^2, that is, $p(\sigma^2) \propto 1/\sigma^2$. The conditional Laplace prior specified in (2.18) applied to linear regression models is known as the *Bayesian lasso*. In their paper, the authors presented two ways of choosing the hyperparameter λ: empirical Bayes through maximum likelikood and full Bayes by using an appropriate hyperprior.

The frequentist ridge regression also has a Bayesian analogue as shown in Equation (2.9) and can also be represented as a SMN hierarchical model. If we let $\lambda = \sigma^2/\psi_j$ in (2.9), we get for each component that $\beta_j \sim N(0, \psi_j)$. Although ψ_j is in principle fixed to achieve equivalence to the classical formulation, even the classical formulation has a data-driven determination of λ by using cross validation. Therefore, in the Bayesian perspective one could also assign a hyperprior for ψ_j, where the most common choice is $\psi_j \sim \mathcal{IG}(a, b)$. It can be shown that assuming the following independent priors

$$\beta_j|\psi_j \overset{\text{ind}}{\sim} \mathcal{N}(0, \psi_j), \qquad \psi_j \sim \mathcal{IG}(a, b), \tag{2.19}$$

for $j = 1, ..., q$, then, β_j will follow a scaled t distribution with $2a$ degrees of freedom and scale parameter $\sqrt{a/b}$, marginally. In that way, the Bayesian version of the ridge regression leads to weaker penalization of large coefficients as long as the t distribution has heavier tails than the Gaussian distribution.

The *elastic net* from Zou and Hastie (2005), which penalty function was showed in Equation (2.13), can also be expressed under the Bayesian perspective. In their paper, it was pointed out that solving the elastic net optimization problem is equivalent to finding the marginal posterior mode of $p(\boldsymbol{\beta}|\boldsymbol{y})$ when the prior distribution is given by

$$p(\boldsymbol{\beta}) \propto \exp\{-\lambda_1 \|\boldsymbol{\beta}\|_1 - \lambda_2 \|\boldsymbol{\beta}\|_2^2\},$$

a compromise between Gaussian and Laplace priors. However, as noted by Li et al. (2010), neither the posterior mode of $p(\boldsymbol{\beta}|\sigma^2, \boldsymbol{y})$ nor the marginal posterior mode of $p(\boldsymbol{\beta}|\boldsymbol{y})$ would be equivalent to the elastic net estimator unless the analysis is conditional on σ^2 or σ^2 is given a point-mass prior. Based on this discussion, Li et al. (2010) proposed a conditional prior specification similar to Park and Casella (2008). The proposed hierarchical prior for $\boldsymbol{\beta}$ is

$$p(\boldsymbol{\beta}|\sigma^2) \propto \exp\left\{-\frac{1}{2\sigma^2}(\lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \|\boldsymbol{\beta}\|_2^2)\right\},$$

while is given an non informative prior for σ^2, i.e., $p(\sigma^2) \propto 1/\sigma^2$.

2.2.2 The Normal-Gamma prior

Even more shrinkage than the Bayesian lasso and the ridge prior can be achieved by using a Gamma mixing distribution in Equation (2.16). The hierarchical formulation, known as the *Normal-Gamma prior* and applied to regression problems in Griffin et al. (2010) is as follows

$$\beta_j|\psi_j \overset{\text{ind}}{\sim} \mathcal{N}(0,\psi_j), \qquad \psi_j|\lambda,\gamma^2 \sim \mathcal{G}(\lambda, 1/(2\gamma^2)), \tag{2.20}$$

for $j = 1, ..., q$, where $\mathcal{G}(\lambda, 1/(2\gamma^2))$ denotes the Gamma distribution with shape parameter λ and mean $2\lambda\gamma^2$. The marginal density $p(\beta_j|\lambda,\gamma^2)$ can be expressed in closed form as follows

$$p(\beta_j) = \frac{1}{\sqrt{\pi}2^{\lambda-1/2}\gamma^{\lambda+1/2}\Gamma(\lambda)}|\beta_j|^{\lambda-1/2}K_{\lambda-1/2}(|\beta_j|/\gamma), \tag{2.21}$$

where K is the modified Bessel function of the third kind. The variance of β_j is $2\lambda\gamma^2$ and the excess kurtosis is $3/\lambda$.

The Gamma distribution can represent a wide-range of shapes. As the shape parameter λ decreases these include distributions that place a lot of mass close to zero but at the same time have heavy tails. The figure 2.1 shows the effect of the shrinking parameter λ on the marginal log prior distribution of β_j. The marginal distribution becomes more peaked at zero which places increasing mass close to zero as λ decreases.

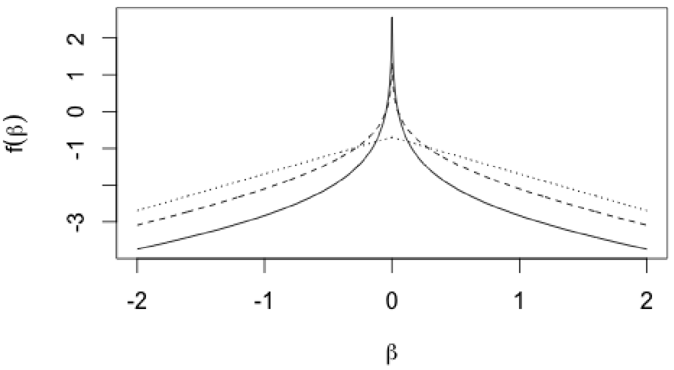


Figure 2.1: *The log density of the normal-gamma prior with a variance of 2 and different values of λ (dot line: $\lambda = 0.1$; dot-dashed line: $\lambda = 0.33$; solid line: $\lambda = 1$)*

The effect of parameter λ of the Normal-Gamma prior is related to shrinking. An interesting interpretation arises when the regressors in $\boldsymbol{X}$ have been standardized so that the mean and variance of each regressor is 0 and 1, respectively. Assuming independent Normal-Gamma priors as in (2.20), the regression total variability is $\sum_{j=1}^{q}\psi_j$. Thus, $\zeta_j = \psi_j/(\sum_{j=1}^{q}\psi_j)$ can be interpreted as the proportion of total variability attributable to the jth regressor and $\zeta = (\zeta_1, ..., \zeta_q)$ follows a Dirichlet distribution with all parameters equal to λ. Therefore, increasing λ will lead to more evenly distributed values of $\zeta = (\zeta_1, ..., \zeta_q)$ and small values of λ will be associated with large differences between the proportions.

Evidently, the Normal-Gamma prior reduces to the Bayesian lasso when $\lambda = 1$. Another interesting case is obtained when $\lambda = 1/2$, in which case $\psi_j|\gamma^2 \sim \gamma^2\chi_1^2$, or equivalently, $\sqrt{\psi_j} \sim N(0,\gamma^2)$. This case motivated the work of Frühwirth-Schnatter and Wagner (2010), which will be discussed in later in Chapter 3.

One could use the empirical Bayes approach for choosing the hyperparameters λ and γ^2.

However, accordingly to Griffin et al. (2010), the posterior distribution of λ and γ^2 can be highly multimodal and an empirical Bayes approach difficult to implement. Therefore, the authors took a fully Bayesian approach assigning hyperpriors to λ and γ^2. A prior which seemed to work well in the simulations is taking λ to be an exponential distribution with mean 1, which offers variability around the Bayesian lasso ($\lambda = 1$).

The prior for the scale parameter γ conditional on λ is defined through a prior on the marginal variance $\mathrm{Var}(\beta_j) = 2\lambda\gamma^2 \sim \mathcal{IG}(2, M)$, so that it has expectation M. This is the same of specifying $\gamma^2|\lambda \sim \mathcal{IG}(2, M/2\lambda)$. The hyperparameter M is chosen to be equal to $M = \frac{1}{q}\sum_{j=1}^{q}\hat{\beta}_j^2$, when the regression design matrix $\boldsymbol{X}$ is non singular, where $\hat{\beta}_j$ is the least square estimate. When $\boldsymbol{X}$ is singular, such as in high-dimensional cases ($q \gg n$), $M = \frac{1}{n}\sum_{j=1}^{q}\tilde{\beta}_j^2$, where $\tilde{\beta}_j$ states for the minimum length least squares estimate. Lastly, they choose a vague prior for the error variance σ^2 so that $\sigma^{-2} \propto 1$.

In summary, the model proposed by Griffin et al. (2010) has the following hierarchical structure

$$
\begin{aligned}
\boldsymbol{y} &\sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}),\\
\boldsymbol{\beta}|\Psi &\sim \mathcal{N}(0, \Psi),\ \Psi = \mathrm{diag}(\psi_1, ..., \psi_q),\\
(\psi_j|\lambda, \gamma) &\sim \mathcal{G}(\lambda, 1/(2\gamma^2)),\\
\gamma^2|\lambda &\sim \mathcal{IG}(2, M/2\lambda),\\
\lambda &\sim \mathcal{E}(1),\\
\sigma^{-2} &\propto 1.
\end{aligned}
$$

The posterior distribution of the parameters can be simulated using a Gibbs sampler with an additional Metropolis-Hastings update. The Gibbs sampler and the respective full conditionals is presented in Algorithm 1.

Note that sampling $\boldsymbol{\beta}$ involves an inversion of an $(q \times q)$ matrix. It is computationally convenient in problems with $q > n$ to express the mean and variance of this distribution using the following form which only involves the inversion of an $(n \times n)$ matrix. We know that in high dimensional problems the standard OLS/MLE estimator is not defined. However, the problem can be re-expressed in terms of a n-dimensional parameter $\boldsymbol{\theta}$ for which the MLE exists. The singular value decomposition of $\boldsymbol{X}$ is

$$\boldsymbol{X} = \boldsymbol{F}'\boldsymbol{D}\boldsymbol{A}',$$

where $\boldsymbol{A}$ is a $(n \times q)$-dimensional matrix such that $\boldsymbol{A}'\boldsymbol{A} = \boldsymbol{I}$, $\boldsymbol{D}$ is a $(n \times n)$-dimensional diagonal matrix and $\boldsymbol{F}$ is a $(n \times n)$-dimensional matrix for which $\boldsymbol{F}'\boldsymbol{F} = \boldsymbol{F}\boldsymbol{F}' = \boldsymbol{I}$. Therefore, we can write

$$\boldsymbol{X}\boldsymbol{\beta} = (\boldsymbol{F}'\boldsymbol{D})\boldsymbol{\theta}.$$

Then, the MLE $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ exists and has the form

$$\hat{\boldsymbol{\theta}} = \boldsymbol{D}^{-1}\boldsymbol{F}\boldsymbol{y}.$$

After some calculation we can express the posterior mean and posterior variance (knowing

Algorithm 1: Gibbs sampler for the Normal-Gamma prior

1. Update $\boldsymbol{\beta}$ by its full conditional

$$(\boldsymbol{\beta}|\boldsymbol{\Psi},\boldsymbol{X},\sigma^2,\boldsymbol{y}) \sim \mathcal{N}(\boldsymbol{X}'\boldsymbol{X}+\sigma^2\Psi^{-1})^{-1}\boldsymbol{X}'\boldsymbol{y},\sigma^2(\boldsymbol{X}'\boldsymbol{X}+\sigma^2\Psi^{-1})^{-1}),$$

where $\boldsymbol{\Psi} = \text{diag}(\psi_1, ..., \psi_q)$.
If response is not centered, in order to allow for an intercept, use $\boldsymbol{X}^* = [1 : \boldsymbol{X}]$ in place of $\boldsymbol{X}$ and $\boldsymbol{\Lambda} = \text{diag}(0, 1/\psi_1, ..., 1/\psi_q)$ in place of $\boldsymbol{\Psi}^{-1}$.

2. Update σ^2 by its full conditional

$$(\sigma^2|\boldsymbol{X},\boldsymbol{y}) \sim \mathcal{IG}(n/2, (\boldsymbol{y}-\boldsymbol{X\beta})'(\boldsymbol{y}-\boldsymbol{X\beta})/2).$$

3. Update $\boldsymbol{\Psi}$ in a block, since $\psi_1, ..., \psi_q$ are independent and each full conditional of ψ_j is

$$\psi_j|\boldsymbol{\beta},\gamma,\lambda \sim \mathcal{GIG}(\lambda-1/2, 1/\gamma^2, \beta_j^2),$$

where $\mathcal{GIG}(m, c, d)$ has the density

$$\frac{(c/d)^{m/2}}{2K_m\sqrt{cd}}x^{m-1}\exp\{-\frac{1}{2}(cx+d/x)\}.$$

An algorithm for simulating the Generalized Inverse Gaussian (GIG) distribution is available in Matlab (randraw toolbox) and in R (package GIGrvg).

4. Update λ by its full conditional

$$\lambda|\gamma^2 \sim \frac{1}{\lambda^2}\exp\left\{-\frac{M}{2\lambda\gamma^2}-\lambda\right\}\left(\prod_{j=1}^{q}\psi_j\right)^{\lambda}\frac{1}{(\Gamma(\lambda))^q(2\gamma^2)^{q\lambda}},$$

which can be updated using a Metropolis-Hastings random walk update on $\log\lambda$. The proposal is $\lambda^* = \exp\{\sigma_\lambda^2 z\}\lambda$, where z is a standard Normal and σ_λ^2 is a tuning parameter which is chosen to set the average acceptance rate at around 20-30%. λ^* is accepted with probability

$$\min\left\{1, \frac{\pi(\lambda^*)}{\pi(\lambda)}\left(\frac{\Gamma(\lambda)}{\Gamma(\lambda^*)}\right)^q\left((2\gamma^2)^{-q}\prod_{j=1}^{q}\psi_j\right)^{\lambda^*-\lambda}\frac{\lambda^*}{\lambda}\right\},$$

where $\pi(\lambda) = (1/\lambda)^2\exp\{-M/(2\gamma^2\lambda)-\lambda\}$.

5. Update γ^2 by its full conditional

$$(\gamma^2|\lambda,\boldsymbol{\Psi}) \sim \mathcal{IG}\left(2+q\lambda, \frac{M}{2\lambda}+\frac{1}{2}\sum_{j=1}^{q}\psi_j\right).$$

that the posterior is Gaussian) as

$$\mathbb{E}(\boldsymbol{\beta}|\boldsymbol{\Psi},\boldsymbol{X},\sigma^2,\boldsymbol{y}) = \boldsymbol{\Psi A}(\boldsymbol{A}'\boldsymbol{\Psi A}+\sigma^2\boldsymbol{D}^{-2})^{-1}\hat{\boldsymbol{\theta}},$$

$$\text{Var}(\boldsymbol{\beta}|\boldsymbol{\Psi},\boldsymbol{X},\sigma^2,\boldsymbol{y}) = \boldsymbol{\Psi} - \boldsymbol{\Psi A}(\boldsymbol{A}'\boldsymbol{\Psi A}+\sigma^2\boldsymbol{D}^{-2})^{-1}\boldsymbol{A}'\boldsymbol{\Psi}.$$

Remember that, we could also fix $\lambda = 1$ to mimic the Bayesian lasso or use a fixed value such as $\lambda = 1/2$ as in Bitto and Frühwirth-Schnatter (2016) or $\lambda = 0.1$ as in Kastner (2016).

The similarity between the Normal-Gamma prior and spike-and-slab prior was discussed in

Griffin et al. (2010). Nevertheless, if the coefficient β_j is small, the shrinkage associated with the spike-and-slab prior tends to be larger than for the matching normal-gamma prior.

2.3 Variable selection

The principle of Occam's Razor states that among several plausible explanations for a phenomenon, the simplest is best. Applied to linear regression analysis, this implies that the smallest model that fits the data is best as unnecessary predictors will add noise to the estimation and, besides that, degrees of freedom will be wasted.

Variable selection is intended to select the best subset of predictors. The problem arises when there is some unknown subset of the predictors with regression coefficients so small that it would be preferable to ignore them.

Classical approaches to variable selection includes stepwise and criterion-based procedures. In stepwise procedure, such as forward selection and backward selection, the choice of predictive variables is carried out by an sequential procedure, where in each step, a variable is considered for addition to or subtraction from the set of explanatory variables while there is a stopping rule based on the value of t-statistic or F-statistic. In criterion-based procedure, if there are q potential predictors, then we fit all 2^q possible models and choose the best one according to some criterion such as AIC, BIC, R^2, Mallow's C_p statistic and others. For a comprehensive summary of these procedures see, e.g, Miller (2002).

Bayesian variable selection is commonly based on spike-and-slab priors for regression coefficients. Consider the Gaussian linear model from Equation (2.1) with q possible predictors and ignoring the intercept (assume that the response is centered). The basic idea is that each component β_j from $\boldsymbol{\beta}$ is modeled as having come either from a distribution with most (or all) of its mass concentrated around zero (the *spike*), or from a comparably diffuse distribution with mass spread out over a large range of values (the *slab*).

By a spike-and-slab model we mean a Bayesian model specified by the following hierarchy

$$\begin{aligned}(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\beta},\sigma^2) &\sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta},\sigma^2\boldsymbol{I}),\\ (\boldsymbol{\beta}|\boldsymbol{\Psi}) &\sim \mathcal{N}(\boldsymbol{0},\boldsymbol{\Psi}),\\ \boldsymbol{\Psi} &\sim \pi(d\boldsymbol{\Psi}),\\ \sigma^2 &\sim \mu(d\sigma^2),\end{aligned} \tag{2.22}$$

where $\boldsymbol{\Psi}$ is a $q \times q$ covariance matrix, π is a prior measure for $\boldsymbol{\Psi}$ and μ is the prior measure for σ^2. Generally, it is assumed that the q coefficients of $\boldsymbol{\beta}$ are independent, in which case the covariance matrix $\boldsymbol{\Psi} = \text{diag}(\psi_1, ..., \psi_q)$.

In the following subsections, we discuss two variable selection approaches that use spike-and-slab models: the *stochastic search variable selection* (SSVS) prior of George and McCulloch (1993) and the Normal mixture of Inverse-Gamma prior of Ishwaran and Rao (2005)), which was originally named *stochastic variable selection* (SVS) by the authors. Then, we generalize the spike-and-slab prior by allowing different combinations of distributions for the spike and the slab as discussed in Frühwirth-Schnatter and Wagner (2011).

2.3.1 The stochastic search variable selection prior

Seminal references about bayesian variable selection using spike-and-slab priors for regression coefficients are Mitchell and Beauchamp (1988) and George and McCulloch (1993). The method developed by the latter, which is called *stochastic search variable selection* (SSVS), is based on embedding the entire regression setup in a hierarchical Bayes normal mixture model, where latent variables are used to identify subset choices. In this framework, the promising subset of predictors can be identified as those with higher posterior probability.

Consider the structure for a spike-and-slab model presented in (2.22). Introducing latent binary variables $J_j \in \{0,1\}$, the prior for the coefficients $\boldsymbol{\beta}$ assumed by George and McCulloch (1993) can be formally expressed as

$$\beta_j | J_j \sim (1 - J_j)\mathcal{N}(0, \tau_j^2) + J_j \mathcal{N}(0, c_j^2 \tau_j^2), \tag{2.23}$$

and

$$p(J_j = 1) = 1 - P(J_j = 0) = \omega_j, \tag{2.24}$$

for $j = 1, .., q$, where $c_j > 1$ is a large scalar and $\tau_j > 0$ is a small scalar. Note that the latent variables $\boldsymbol{J} = (J_1, ..., J_q)$ can assume 2^q values as $\boldsymbol{J} \in \{0,1\}^q$.

The hyperparameter ω_j in (2.24) is the prior probability that the covariate $\boldsymbol{X}_j$ has a significant effect and should be included in the model. A simple and usual choice for ω_j is simply to assume $\omega_j = \omega = 0.5$ for all $j = 1, ..., q$. A more flexible approach, assuming that $\omega_j = \omega$ for all $j = 1, ..., q$, is to place a hyperprior on ω, for instance, a Beta prior is a reasonable and convenient choice, including the special case $\omega \sim \mathcal{U}(0,1)$, where $\mathcal{U}$ denotes the Uniform distribution.

Thus, the prior for the coefficients stated in equations (2.23) and (2.24) is a mixture of Gaussian densities. When $J_j = 0$, we have that $(\beta_j | J_j = 0) \sim \mathcal{N}(0, \tau_j^2)$, and when $J_j = 1$, we have that $(\beta_j | J_j = 1) \sim \mathcal{N}(0, c_j^2 \tau_j^2)$. The interpretation is as follows. First, by setting τ_j small enough, if $J_j = 0$, then β_j is probably so small that it can be safely estimated as being equal to 0. Second, by setting c_j large, if $J_j = 1$, then a non zero estimate of β_j is probably included in the final model.

This mixture prior for each component of $\boldsymbol{\beta}$ can be obtained using a multivariate normal mixture prior

$$\boldsymbol{\beta} | \boldsymbol{J} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{D}_J \boldsymbol{R} \boldsymbol{D}_J), \tag{2.25}$$

where $\boldsymbol{J} = (J_1, ..., J_q)$, $\boldsymbol{R}$ is the prior correlation matrix, and $\boldsymbol{D}_J = \text{diag}(a_1\tau_1, ..., a_q\tau_q)$, with $a_j = 1$ if $J_j = 0$ and $a_j = c_j$ if $J_j = 1$ for each $j = 1, ..., q$. That is, the covariance matrix $\boldsymbol{\Psi}$ from the spike-and-slab model stated in (2.22) can be decomposed using the variance-correlation decomposition, i.e., $\boldsymbol{\Psi} = \boldsymbol{D}_J \boldsymbol{R} \boldsymbol{D}_J$. In general, most researchers work with independent priors for the coefficients, in which case $\boldsymbol{R} = \boldsymbol{I}$.

For the indicators $\boldsymbol{J}$ a standard joint prior was assumed, that is, each component J_j is independent from each other and follows a Bernoulli distribution

$$\boldsymbol{J} | \boldsymbol{\omega} \sim \prod_{j=1}^{q} \omega_j^{J_j} (1 - \omega_j)^{1 - J_j}. \tag{2.26}$$

Finally, it was proposed an Inverse-Gamma prior for the variance σ^2 as

$$\sigma^2|\boldsymbol{J} \sim \mathcal{IG}(\nu_J/2, \nu_J\lambda_J/2), \tag{2.27}$$

where ν_J and λ_J may depend on the indicators variables $\boldsymbol{J}$ to incorporate dependence between $\boldsymbol{\beta}$ and σ^2 as one might expect that σ^2 would decrease as the dimension of $\boldsymbol{\beta}$ increased.

One may ask how to set the hyperparameters c_j and τ_j from (2.23). To respond this, we refer to George and McCulloch (1997). In this second article, they used a simpler notation as follows

$$\beta_j|J_j \sim (1-J_j)\mathcal{N}(0, \upsilon_{0j}) + J_j\mathcal{N}(0, \upsilon_{1j}), \tag{2.28}$$

for $j = 1, ..., q$, where $\upsilon_{0j} = \tau_j^2$ and $\upsilon_{1j} = c_j^2\tau_j^2$.

As discussed in George and McCulloch (1993) and in George and McCulloch (1997), to use the above hierarchical mixture setup for variable selection, the hyperparameters υ_{0j} and υ_{1j} are set small and large ($\upsilon_{0j} \ll \upsilon_{1j}$) or, as mentioned before, $c_j > 1$ is large and $\tau_j > 0$ is small. To help guide the choice of c_j, George and McCulloch (1993) observed that densities of $\mathcal{N}(0, \tau_j^2)$ and $\mathcal{N}(0, c_j^2\tau_j^2)$ intersect at $\xi(c_j)\tau_j$ with

$$\xi(c_j) = \sqrt{2\log(c_j)c_j^2/(c_j^2 - 1)},$$

which implies that the density of $\mathcal{N}(0, c_j^2\tau_j^2)$ is greater than the density of $\mathcal{N}(0, \tau_j^2)$ if, and only if, $|\beta_j| > \xi(c_j)\tau_j$. This property is equivalent to setting any υ_{0j} and υ_{1j} from equation (2.28) satisfying

$$\delta_j^2 = \log(\upsilon_{1j}/\upsilon_{0j})/(\upsilon_{0j}^{-1} - \upsilon_{1j}^{-1}),$$

where $\delta_j = \xi(c_j)\tau_j$. In this case, if $|\beta_j| < \delta_j$, then it would be preferable to exclude covariate $\boldsymbol{X}_j$. Thus, $\delta_j > 0$ could be considered the *threshold of practical significance*. A simple choice for δ_j might be $\Delta\boldsymbol{y}/\Delta\boldsymbol{X}_j$, where $\Delta\boldsymbol{y}$ is the size of an insignificant change in $\boldsymbol{y}$, and $\Delta\boldsymbol{X}_j$ is the size of the maximum feasible change in $\boldsymbol{X}_j$. Note that, the intersection point from the two densities increases very slowly with c_j. For example, the choices $c_j = 10, 100, 1000, 10000, 100000$ correspond to $\xi(c_j) = 2.1, 3.1, 3.7, 4.3, 4.8$.

In addition, George and McCulloch (1997) noted that the incorporation of a threshold δ_j requires choosing $\upsilon_{0j} > 0$ for all j as δ_j depends on the ratio $\upsilon_{1j}/\upsilon_{0j}$. Computational problems can arise when this ratio is set too large, accordingly to the authors. However, this problems should be avoided whenever $\upsilon_{1j}/\upsilon_{0j} \leq 10000$, thus allowing for a wide variety of settings.

Besides using the the threshold of practical significance, an alternative semiautomatic approach for choosing τ_j^2 and c_j based on statistical significance is described in George and McCulloch (1993).

Gibbs sampling the best subsets. In the hierarchical mixture model defined by the equation (2.25), the latent vector $\boldsymbol{J} = (J_1, ..., J_q)$ contains the relevant information for variable selection. If $\boldsymbol{J}$ were known, then with high probability for suitably chosen $\tau_1^2, ..., \tau_q^2$ and $c_1, ..., c_q$, a desirable model would be obtained by including $\boldsymbol{X}_j$ for which $J_j = 1$ and excluding those covariates for which $J_j = 0$.

Although $\boldsymbol{J}$ in unknown, the posterior distribution $p(\boldsymbol{J}|\boldsymbol{y})$ can provide useful information. The posterior updates the prior probabilities on each 2^q possible values of $\boldsymbol{J}$. Thus, identifying those $\boldsymbol{J}$ with higher posterior probability, one can identify the more promising subsets of variables. However, rather than calculate all 2^q posterior probabilities in $p(\boldsymbol{J}|\boldsymbol{y})$, we can use Gibbs sampler to generate a sequence

$$\boldsymbol{J}^{(1)}, ..., \boldsymbol{J}^{(m)},$$

which in many cases converges rapidly in distribution to the posterior $p(\boldsymbol{J}|\boldsymbol{y})$, where m is the number of iterations of the Markov chain. The sequence is embedded in the chain

$$\boldsymbol{\beta}^{(1)}, \sigma^{2(1)}, \boldsymbol{J}^{(1)}, ..., \boldsymbol{\beta}^{(m)}, \sigma^{2(m)}, \boldsymbol{J}^{(m)},$$

which is obtained by successive simulation from the full conditionals $p(\boldsymbol{\beta}|\sigma^2, \boldsymbol{J}, \boldsymbol{y})$, $p(\sigma^2|\boldsymbol{\beta}, \boldsymbol{y})$ and $p(J_j|\boldsymbol{\beta}, \sigma^2, \boldsymbol{J}_{(-j)})$ for $j = 1, ..., q$, where $\boldsymbol{J}_{(-j)} = (J_1, ..., J_{j-1}, J_{j+1}, ..., J_q)$.

Note that the last full conditional does not depend on $\boldsymbol{y}$, which reduces computational requirements and allows for faster convergence of the sequence $\boldsymbol{J}^{(1)}, ..., \boldsymbol{J}^{(m)}$. In problems where the number of predictors q is small, this sequence can be used to evaluate the entire posterior $p(\boldsymbol{J}|\boldsymbol{y})$. In large problems, the sequence may still provide useful information. This is because those $\boldsymbol{J}$ with highest probability will also appear most frequently in the generated sequence and hence will be easiest to identify.

The Gibbs sampler and the respective full conditionals $p(\boldsymbol{\beta}|\sigma^2, \boldsymbol{J}, \boldsymbol{y})$ and $p(\sigma^2|\boldsymbol{\beta}, \boldsymbol{y})$ for the SVSS model are presented in Algorithm 2.

Algorithm 2: Gibbs sample the SVSS prior

1. Draw $\boldsymbol{\beta}$ from its full conditional

$$(\boldsymbol{\beta}|\sigma^2, \boldsymbol{J}, \boldsymbol{y}) \sim \mathcal{N}((\boldsymbol{X}'\boldsymbol{X} + \sigma^2(\boldsymbol{D}_J\boldsymbol{R}\boldsymbol{D}_J)^{-1})^{-1}\boldsymbol{X}'\boldsymbol{y}, \sigma^2(\boldsymbol{X}'\boldsymbol{X} + \sigma^2(\boldsymbol{D}_J\boldsymbol{R}\boldsymbol{D}_J)^{-1})^{-1}).$$

2. Draw σ^2 from its full conditional

$$(\sigma^2|\boldsymbol{\beta}, \boldsymbol{J}, \boldsymbol{y}) \sim \mathcal{IG}\left(\frac{n + \nu_J}{2}, \frac{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \nu_J\lambda_J}{2}\right).$$

3. The vector $\boldsymbol{J}$ is obtained componentwise by sampling each J_j consecutively from the Bernoulli distribution with probability

$$p(J_j = 1|\boldsymbol{\beta}, \sigma^2, \boldsymbol{J}_{(-j)}) = \frac{a}{a + b},$$

where

$$a = p(\boldsymbol{\beta}|J_j = 1, \boldsymbol{J}_{(-j)})p(\sigma^2|J_j = 1, \boldsymbol{J}_{(-j)})p(\boldsymbol{J}_{(-j)}, J_j = 1),$$
$$b = p(\boldsymbol{\beta}|J_j = 0, \boldsymbol{J}_{(-j)})p(\sigma^2|J_j = 0, \boldsymbol{J}_{(-j)})p(\boldsymbol{J}_{(-j)}, J_j = 0).$$

It is worth pointing out that, under the independent Bernouilli prior for $\boldsymbol{J}$ as in equation (2.26), when the hyperparameters of the prior for σ^2 in equation (2.27) are constant (i.e, $\nu_J = \nu$ and $\lambda_J = \lambda$), and under the choice of independent priors ($\boldsymbol{R} = \boldsymbol{I}$), a and b from the full conditional of $\boldsymbol{J}$ can be obtained more simply by

$$\begin{aligned} a &= p(\beta_j|J_j = 1)\omega_j, \\ b &= p(\beta_j|J_j = 0)(1-\omega_j). \end{aligned}$$

2.3.2 The Normal mixture of Inverse-Gamma prior

Note that the SSVS prior under the choice of independent priors for β_j can be rewritten as

$$\begin{aligned} \beta|\psi_j &\overset{\text{ind}}{\sim} \mathcal{N}(0,\psi_j), \\ (\psi_j|c_j,\tau_j,J_j) &\sim (1-J_j)\delta_{\tau_j^2}(.) + J_j\delta_{c_j^2\tau_j^2}(.), \\ J_j|\omega_j &\sim (1-\omega_j)\delta_0(.) + \omega_j\delta_1(.),\; j=1,...,q, \end{aligned} \tag{2.29}$$

where $\delta_\upsilon(.)$ is a discrete measure concentrated at value υ.

In practice, it can be difficult to select values τ_j^2, $c_j^2\tau_j^2$ and ω_j used in the SSVS prior. Recognizing this problem Ishwaran and Rao (2005) proposed a continuous bimodal distribution for ψ_j in place of the two point mixture distribution for ψ_j in Equation (2.29).

Instead of encouraging variable selection by placing priors directly on the coefficients β_j, for each variable $j = 1,...,q$, a different strategy is to place priors on the variances ψ_j. The hierarchical prior is formalized as

$$\begin{aligned} (\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\beta},\sigma^2) &\sim \mathcal{N}(\boldsymbol{X\beta},\sigma^2\lambda_n\boldsymbol{I}), \\ (\beta_j|\tau_j^2,K_j) &\overset{\text{ind}}{\sim} \mathcal{N}(0,K_j\tau_j^2), \\ K_j|\omega &\overset{\text{iid}}{\sim} (1-\omega)\delta_{\upsilon_0}(.) + \omega\delta_{\upsilon_1}(.), \\ \tau_j^{-2} &\overset{\text{iid}}{\sim} \mathcal{G}(a_\tau,b_\tau), \\ \omega &\sim \mathcal{B}(a_\omega,b_\omega), \\ \sigma^{-2} &\sim \mathcal{G}(a_\sigma,b_\sigma), \end{aligned} \tag{2.30}$$

for $j = 1,...,q$, where $\mathcal{G}$ denotes the Gamma distribution with shape a_τ and rate b_τ (mean a_τ/b_τ), and $\mathcal{B}$ denotes the Beta distribution with parameters a_ω and b_ω.

Note that each variance $\psi_j = K_j\tau_j^2$, where $K_j \in \{\upsilon_0,\upsilon_1\}$. The prior for β_j in (2.30) can also be rewritten as

$$\begin{aligned} \beta_j|\psi_j &\overset{\text{ind}}{\sim} \mathcal{N}(0,\psi_j), \\ \psi_j|I_j &\sim (1-J_j)\mathcal{IG}(a_\tau,\upsilon_0 b_\tau) + J_j\mathcal{IG}(a_\tau,\upsilon_1 b_\tau), \end{aligned} \tag{2.31}$$

for $j = 1,...,q$, where $J_j \in \{0,1\}$ is related to the binary variable $K_j \in \{\upsilon_0,\upsilon_1\}$ in the original formulation (Equation (2.30)) through $J_j = 1 \Leftrightarrow K_j = \upsilon_1$ and $J_j = 0 \Leftrightarrow K_j = \upsilon_0$. Thus, we see that this structure means a spike-and-slab prior for variances specified as a bimodal mixture of two Inverse-Gamma distributions, also known as Normal mixture of Inverse-Gamma (NMIG) prior.

The NMIG prior provides a natural procedure for selecting variables. The larger the estimated posterior probability of the binary variable K_j, i.e., the higher is the percentage of υ_1 values in the sample, the larger is the evidence that the jth covariate has non-negligible effects and can not be eliminated from the regression model.

Rescaled spike-and-slab models. Ishwaran and Rao (2005) suggested default options, in particular $\upsilon_1 = 1$, after standardizing all covariates and rescaling $\boldsymbol{y}$. They have called this specific approach *rescaled spike-and-slab models*, that is, in place of $\boldsymbol{y}$ they used

$$y_i^* = \hat{\sigma}_n^{-1} n^{1/2} y_i, \qquad \hat{\sigma}_n^2 = \left\| \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} \right\|^2 / (n-q),$$

for $i = 1, ..., n$, where $\hat{\boldsymbol{\beta}}$ is the least square estimate for $\boldsymbol{\beta}$ and $\hat{\sigma}_n^2$ is the unbiased estimate for σ^2. The *rescaled spike-and-slab model* is the same from Equation (2.30), except from $\boldsymbol{y}$ and υ_1. While $\boldsymbol{y}$ is substituted by

$$\boldsymbol{y}^* \sim \mathcal{N}(\boldsymbol{X\beta}, \sigma^2 \lambda_n \boldsymbol{I}), \tag{2.32}$$

where λ_n is a variance inflation factor introduced to compensate for the scaling of $\boldsymbol{y}$, a fixed value was used for $\upsilon_1 = 1$ in (2.30) so that

$$K_j | \omega \overset{\text{iid}}{\sim} (1-\omega)\delta_{\upsilon_0}(.) + \omega \delta_1(.). \tag{2.33}$$

Assuming $\upsilon_1 = 1$ and integrating out ψ_j from Equation (2.31) , the marginal distribution $\beta_j|\omega$ is a mixture of two scaled t-distributions:

$$\begin{aligned}
p(\beta_j|\omega) &= \omega \int_0^\infty p(\beta_j|\psi_j) p(\psi_j|a_\tau, b_\tau) d\psi_j + (1-\omega) \int_0^\infty p(\beta_j|\psi_j) p(\psi_j|a_\tau, b_\tau \upsilon_0) d\psi_j \\
&= K_1 \int_0^\infty \psi_j^{-(a_\tau+3/2)} \exp\left\{ -\frac{\frac{\beta_j^2}{2} + b_\tau}{\psi_j} \right\} d\psi_j + K_2 \int_0^\infty \psi_j^{-(a_\tau+3/2)} \exp\left\{ -\frac{\frac{\beta_j^2}{2} + b_\tau \upsilon_0}{\psi_j} \right\} d\psi_j \\
&= K_1 \int_0^\infty \left(\frac{\beta_j^2}{2} + b_\tau \right)^{-(a_\tau+1/2)} \left(\frac{\psi}{\frac{\beta_j^2}{2} + b_\tau} \right)^{-(a_\tau+3/2)} \exp\left\{ -\frac{\frac{\beta_j^2}{2} + b_\tau}{\psi_j} \right\} d\left(\frac{\psi}{\frac{\beta_j^2}{2} + b_\tau} \right) \\
&+ K_2 \int_0^\infty \left(\frac{\beta_j^2}{2} + b_\tau \upsilon_0 \right)^{-(a_\tau+1/2)} \left(\frac{\psi}{\frac{\beta_j^2}{2} + b_\tau \upsilon_0} \right)^{-(a_\tau+3/2)} \exp\left\{ -\frac{\frac{\beta_j^2}{2} + b_\tau \upsilon_0}{\psi_j} \right\} d\left(\frac{\psi}{\frac{\beta_j^2}{2} + b_\tau \upsilon_0} \right) \\
&= \omega \frac{b_\tau^{a_\tau} \Gamma(a_\tau + 1/2)}{\sqrt{2\pi}\Gamma(a_\tau) \left(\frac{\beta_j^2}{2} + b_\tau \right)^{(a_\tau+1/2)}} + (1-\omega) \frac{(b_\tau \upsilon_0)^{a_\tau} \Gamma(a_\tau + 1/2)}{\sqrt{2\pi}\Gamma(a_\tau) \left(\frac{\beta_j^2}{2} + b_\tau \upsilon_0 \right)^{(a_\tau+1/2)}} \\
&= \omega \frac{\Gamma(\frac{2a_\tau+1}{2})}{\Gamma(\frac{2a_\tau}{2}) \sqrt{2\pi a_\tau \frac{b_\tau}{a_\tau}}} \left(1 + \frac{\beta_j^2}{2a_\tau \frac{b_\tau}{a_\tau}} \right)^{-(\frac{2a_\tau+1}{2})} + (1-\omega) \frac{\Gamma(\frac{2a_\tau+1}{2})}{\Gamma(\frac{2a_\tau}{2}) \sqrt{2\pi a_\tau \frac{b_\tau \upsilon_0}{a_\tau}}} \left(1 + \frac{\beta_j^2}{2a_\tau \frac{b_\tau \upsilon_0}{a_\tau}} \right)^{-(\frac{2a_\tau+1}{2})},
\end{aligned}$$

which is a mixture of scaled t-distributions with $2a_\tau$ degrees of freedom, scale factors $\sqrt{b_\tau/a_\tau}$ and $\sqrt{(b_\tau \upsilon_0)/a_\tau}$ and weights ω and $1-\omega$, where we used the fact that $p(J_j = 1) = P(K_j = \upsilon_1) = \omega$ and that $K_1 = \omega \frac{b_\tau^{a_\tau}}{\sqrt{2\pi}\Gamma(a_\tau)}$ and $K_2 = (1-\omega) \dfrac{(b_\tau \upsilon_0)^{a_\tau}}{\sqrt{2\pi}\Gamma(a_\tau)}$.

It was shown in Ishwaran and Rao (2005) that, under some conditions for design matrix $\boldsymbol{X}$, the parameter λ_n in Equation (2.32) controls the amount of shrinkage and that a value of $\lambda_n = n$ is the amount of penalization required in order to ensure shrinkage effect in the limit.

Let $\hat{\boldsymbol{\beta}}_n^*(\Psi, \sigma^2) = \mathbf{E}(\boldsymbol{\beta}|\boldsymbol{\Psi}, \sigma^2, \boldsymbol{y}^*)$ be the conditional posterior mean for $\boldsymbol{\beta}$ from the *rescaled spike-and-slab model*, where $\boldsymbol{\Psi} = \text{diag}(K_1\tau_1^2, ..., K_q\tau_q^2)$. It can be verified that

$$\begin{aligned}\hat{\boldsymbol{\beta}}_n^*(\boldsymbol{\Psi},\sigma^2) &= (\boldsymbol{X}'\boldsymbol{X}+\sigma^2\lambda_n\boldsymbol{\Psi}^{-1})^{-1}\boldsymbol{X}'\boldsymbol{y}^* \\ &= \hat{\sigma}_n^{-1}n^{1/2}(\boldsymbol{X}'\boldsymbol{X}+\sigma^2\lambda_n\boldsymbol{\Psi}^{-1})^{-1}\boldsymbol{X}'\boldsymbol{y}.\end{aligned}$$

Thus, by the ridge prior presented in Equation (2.9), we find that $\hat{\boldsymbol{\beta}}_n^*(\boldsymbol{\Psi},\sigma^2)$ is the ridge solution to a regression of $\boldsymbol{y}^*$ on $\boldsymbol{X}$ with ridge matrix $\sigma^2\lambda_n\boldsymbol{\Psi}^{-1}$.

Now define $\hat{\boldsymbol{\theta}}_n^*(\boldsymbol{\Psi},\sigma^2) = \hat{\sigma}_n\hat{\boldsymbol{\beta}}_n^*(\boldsymbol{\Psi},\sigma^2)/\sqrt{n}$. Then, it is clear that $\hat{\boldsymbol{\theta}}_n^*(\boldsymbol{\Psi},\sigma^2)$ is the ridge solution to a regression of $\boldsymbol{y}$ on $\boldsymbol{X}$ with ridge matrix $\sigma^2\lambda_n\boldsymbol{\Psi}^{-1}$. In that way, λ_n can be seen as a penalty term because ridge solution can always be recast as an optimization problem, that is, it is straightforward to prove that

$$\hat{\boldsymbol{\theta}}_n^*(\boldsymbol{\Psi},\sigma^2) = \underset{\boldsymbol{\beta}}{\arg\min}\left\{\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|^2+\lambda_n\sum_{j=1}^{q}\sigma^2\psi_j^{-1}\beta_j^2\right\}.$$

Theorem 2 from Ishwaran and Rao (2005) is an analogue of Theorem 1 from Knight and Fu (2000), which establishes consistency for bridge estimators, where ridge estimation is a special case. Define $\hat{\boldsymbol{\beta}}_n^* = \mathbf{E}(\boldsymbol{\beta}|\boldsymbol{y}^*)$ and $\hat{\boldsymbol{\theta}}_n^* = \hat{\sigma}_n\hat{\boldsymbol{\beta}}_n^*/\sqrt{n}$. Through the mentioned theorem, the authors demonstrated that a penalization satisfying $\lambda_n/n \to 0$ yields a posterior mean after rescaling $\hat{\boldsymbol{\theta}}_n = \hat{\sigma}_n\hat{\boldsymbol{\beta}}_n/\sqrt{n} \xrightarrow{p} \boldsymbol{\beta}$, assuming that $\boldsymbol{X}'\boldsymbol{X}$ is positive definite and that $\boldsymbol{X}'\boldsymbol{X}/n \to \boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ is positive definite. That is, if $\lambda_n/n \to 0$, the posterior mean after rescaling is asymptotically consistent for $\boldsymbol{\beta}$.

While consistency is crucial for estimation purposes, it could be advantageous in terms of variable selection to have a shrinkage effect that does not vanish asymptotically and a posterior mean that behaves differently from OLS. For this result the authors assumed $\lambda_n = n$. When this is assumed, it was found that σ^2 plays an important adaptive role in adjusting the penalty λ_n. It was also noted that under this setting the posterior of σ^2 would concentrate around the value of 1.

The rescaled spike-and-slab model uses a Gibbs sampler to simulate the posterior distribution $p(\boldsymbol{\beta},\boldsymbol{K},\boldsymbol{\tau}^2,\omega,\sigma^2|y^*)$. Recall that $\psi_j = K_j\tau_j^2$, $j = 1,...,q$, so simulating $\boldsymbol{K} = (K_1,...,K_q)$ and $\boldsymbol{\tau}^2 = (\tau_1^2,...,\tau_q^2)$ provides an update for $\boldsymbol{\Psi} = \text{diag}(\psi_1,...,\psi_q)$. The sampler is presented below in Algorithm 3.

2.3.3 Other mixture priors

Although the NMIG prior specified by (2.30) allows discrimination or variable selection, it does not encourage shrinkage in the sense that the resulting marginal distribution of each coefficient β_j is a two component mixture of scaled Student's t distributions as shown before through the Section 2.3.2. That is, assuming the NMIG structure as presented before, each β_j has the following marginal distribution

$$\beta_j|\omega \sim \omega t_{2a_\tau}(0,\upsilon_1 b_\tau/a_\tau)+(1-\omega)t_{2a_\tau}(0,\upsilon_0 b_\tau/a_\tau),$$

where $t_\xi(0,s)$ denotes the Student's t distribution with zero location, scale $\sqrt{s}$ and ξ degrees of freedom. Note that the marginal distribution of β_j is also a spike-and-slab prior.

Algorithm 3: Gibbs sampler for the NMIG prior

1. Simulate $\boldsymbol{\beta}$ from its full conditional

$$(\boldsymbol{\beta}|\boldsymbol{\Psi}, \sigma^2, \boldsymbol{y}^*) \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2\boldsymbol{\Sigma}),$$

where

$$\boldsymbol{\mu} = \boldsymbol{\Sigma}\boldsymbol{X}'\boldsymbol{y}^*, \qquad \boldsymbol{\Sigma} = (\boldsymbol{X}'\boldsymbol{X} + \sigma^2 n\boldsymbol{\Psi}^{-1})^{-1}.$$

For large q the inversion from the first step can be very costly. A better approach is to update $\boldsymbol{\beta}$ individually

$$(\beta_j|\psi_j, \sigma^2, \boldsymbol{y}^*) \sim \mathcal{N}(\mu_j, \Sigma_j),$$

with

$$\mu_j = \Sigma_j\boldsymbol{X}_j'(\boldsymbol{y}^* - \boldsymbol{X}_{(-j)}\boldsymbol{\beta}_{(-j)}), \qquad \Sigma_j = (\boldsymbol{X}_j'\boldsymbol{X}_j + \sigma^2 n\psi_j^{-1})^{-1},$$

where the subscript $(-j)$ denotes all the covariates except $\boldsymbol{X}_j$.

2. Simulate $\boldsymbol{K}$ from its full conditional

$$(K_j|\boldsymbol{\beta}, \boldsymbol{\tau}^2, \omega) \overset{\text{ind}}{\sim} \frac{\omega_{1j}}{\omega_{1j}+\omega_{2j}}\delta_{\upsilon_0}(.) + \frac{\omega_{2j}}{\omega_{1j}+\omega_{2j}}\delta_1(.),\ j = 1, ..., q,$$

where

$$\omega_{1j} = (1-\omega)\upsilon_0^{-1/2}\exp\left(-\frac{\beta_j^2}{2\upsilon_0\tau_j^2}\right), \qquad \omega_{2j} = \omega\exp\left(-\frac{\beta_j^2}{2\tau_j^2}\right).$$

3. Simulate τ_j^2 from its full conditional

$$(\tau_j^2|\boldsymbol{\beta}, \boldsymbol{K}) \overset{\text{ind}}{\sim} \mathcal{IG}\left(a_\tau + \frac{1}{2} + b_\tau\frac{\beta_j^2}{2K_j}\right),\ j = 1, ..., q.$$

4. Simulate ω from its full conditional

$$(\omega|K_j) \sim \mathcal{B}(a_\omega + \#\{j : K_j = 1\}, b_\omega + \#\{j : K_j = \upsilon_0\}).$$

5. Simulate σ^2 from its full conditional

$$\sigma^2|\boldsymbol{\beta}, \boldsymbol{y}^* \sim \mathcal{IG}\left(a_\sigma + \frac{n}{2}, b_\sigma + \frac{1}{2n}\|\boldsymbol{y}^* - \boldsymbol{X}\boldsymbol{\beta}\|^2\right).$$

6. Update $\boldsymbol{\Psi}$ by setting $\psi_j = K_j\tau_j^2$ for $j = 1, ..., q.$ and repeat the previous steps for a sufficiently large number of iterations until the chain converges.

Hence it makes sense to choose other component specific distributions, besides the Inverse Gamma, that could actually induce shrinkage. Consider the binary variable $J_j \in \{0, 1\}$ which indicates the spike ($J_j = 0$) or the slab ($J_j = 1$). Assume that each β_j has a spike-and-slab mixture distribution as

$$\beta_j|J_j, \boldsymbol{\theta} \sim J_j p_{slab}(\beta_j|\boldsymbol{\theta}) + (1 - J_j)p_{spike}(\beta_j|\boldsymbol{\theta}), \tag{2.34}$$

where $\boldsymbol{\theta}$ is a vector of parameters, $p_{spike}(\beta_j|\boldsymbol{\theta})$ is the spike distribution and $p_{slab}(\beta_j|\boldsymbol{\theta})$ is the slab distribution.

One can achieve (2.34) by formulating appropriate spike-and-slab priors to the component variances $\operatorname{Var}_{spike}(\beta_j|\boldsymbol{\theta})$ and $\operatorname{Var}_{slab}(\beta_j|\boldsymbol{\theta})$, similar to what was proposed by Ishwaran and Rao (2005). An excellent reference that discusses different combinations of spike-and-slab distributions is Frühwirth-Schnatter and Wagner (2011). Following their notation, we note that each coefficient β_j can be defined hierarchically as a scaled mixture of Normals (SMN), i.e., $\beta_j|\psi_j \sim \mathcal{N}(0, \psi_j), \quad \psi_j|\theta \sim p(\psi_j|\boldsymbol{\theta})$.

Therefore, if we assume absolutely continuous priors for $\psi_j|J_j = 1$ (slab) and $\psi_j|J_j = 0$ (spike), that is, a mixture prior for ψ_j, then we reach the spike-and-slab structure for the coefficient β_j as in (2.34). For instance, choosing Inverted-Gamma densities for both the spike and the slab variances can be done by

$$(\psi_j|J_j = 0) \sim \mathcal{IG}(\nu, rQ), \qquad (\psi_j|J_j = 1) \sim \mathcal{IG}(\nu, Q), \tag{2.35}$$

where $\boldsymbol{\theta} = (\nu, r, Q)$ is the vector of hyperparameters that define the conditional (on I_j) variances' densities. Note that (2.35) is the same NMIG prior presented in Section (2.30), assuming that $a_{\mathcal{T}} = \nu$, $b_{\mathcal{T}} = Q$, $r = \upsilon_0$ and $\upsilon_1 = 1$. Thus, the hyperparameter r can be viewed as the ratio of the variances $r = \operatorname{Var}_{spike}(\beta_j|\boldsymbol{\theta})/\operatorname{Var}_{slab}(\beta_j|\boldsymbol{\theta})$, which we have shown to be a small number less than 1.

Another mixture prior for ψ_j arises by choosing Exponential densities for both the spike and the slab as

$$(\psi_j|J_j = 0) \sim \mathcal{E}(1/2rQ), \qquad (\psi_j|J_j = 1) \sim \mathcal{E}(1/2Q), \tag{2.36}$$

where $\mathcal{E}(\alpha)$ denotes the Exponential distribution with mean $1/\alpha$, which leads to a mixture of Laplace densities for β_j as noted by Frühwirth-Schnatter and Wagner (2011), that is

$$(\beta_j|\omega) \sim \omega Lap(\sqrt{Q}) + (1-\omega) Lap(\sqrt{rQ}), \tag{2.37}$$

where $Lap(x)$ denotes the Laplace distribution with mean 0 and scale parameter x and the weight ω is the prior probability of the slab, i.e., $\omega = p(J_j = 1)$. If we assume that ψ_j is a mixture of Gammas, that is,

$$(\psi_j|J_j = 0) \sim \mathcal{G}(a, 1/2rQ), \qquad (\psi_j|J_j = 0) \sim \mathcal{G}(a, 1/2Q), \tag{2.38}$$

then the marginal distribution of β_j is a mixture of Normal-Gamma distributions as discussed in Section 2.2.2

$$(\beta_j|\omega) \sim \omega \mathcal{NG}(\beta_j|a, Q) + (1-\omega)\mathcal{NG}(\beta_j|a, r, Q).$$

More generally we may combine distribution families which lead to shrinkage for the spike and, at the same time, avoid too much smoothing in the slab of the marginal mixture of β_j. A promising candidate, which was used by Frühwirth-Schnatter and Wagner (2011), is combining the Exponential density for the spike with the Inverted Gamma density for the slab as follows

$$(\psi_j|J_j = 0) \sim \mathcal{E}(1/2rQ), \qquad (\psi_j|J_j = 1) \sim \mathcal{IG}(\nu, Q), \tag{2.39}$$

which leads to a finite mixture for β_j, where a Laplace density in the spike is combined with a Student-t distribution in the slab as

$$\beta_j|\omega \sim \omega t_{2\nu}(0, Q/\nu) + (1-\omega)Lap(\sqrt{rQ}). \tag{2.40}$$

From what was exposed above, we conclude that the variances of the slab and the spike depend on the hyperparameter Q. One could simply fix Q or place and hyperprior on it, which was the strategy adopted by Frühwirth-Schnatter and Wagner (2011). Assuming an Inverted-Gamma prior for the variance of β_j leads to a general prior distribution for $Q|\omega$ as follows

$$v_\beta = \text{Var}(\beta_j|Q,\omega) = (1-\omega)\,\text{Var}_{spike}(\beta_j|r,Q) + \omega\,\text{Var}_{slab}(\beta_j|Q) \sim \mathcal{IG}(c_0, C_0), \tag{2.41}$$

where r is a fixed parameter. Thus, the distribution of $Q|\omega$ is

$$Q|\omega \sim \mathcal{IG}(c_0, C_0/s^*(\omega)), \tag{2.42}$$

where $s^*(\omega)$ is a function of ω and other hyperparameters, depending on the mixing distribution adopted.

If we assume the same mixing distribution for both the spike and the slab, then we can conclude that $\text{Var}_{spike}(\beta_j|r,Q) = cQr$ and $\text{Var}_{slab}(\beta_j|Q) = cQ$, where c is a constant that depends on the distribution assumption. Thus, $s^*(\omega)$ from Equation (2.42) turns into $s^*(\omega) = c[(1-\omega)r + \omega]$.

For the NMIG prior (whose marginal distribution for β_j is scaled-t) we have that

$$\text{Var}_{slab}(\beta_j|Q) = \frac{2\nu}{2\nu-2}\frac{Q}{\nu},$$

thus, the constant $c = 1/(\nu-1)$. For the mixture of Normal-Gammas, we have

$$\text{Var}_{slab}(\beta_j|Q) = 2aQ,$$

then, $c = 2a$ depends on the shape parameter a specified in (2.38). For the mixture of Laplaces, we have

$$\text{Var}_{slab}(\beta_j|Q) = 2Q,$$

then, $c = 2$. Finally, if we adopt the prior in (2.40), then assuming that $\text{Var}_{spike}(\beta_j|r,Q) = c_1Qr$ and $\text{Var}_{slab}(\beta_j|r,Q) = c_2Q$ and using the results from the variances of Laplace and t-distribution, we have that $c_1 = 2$ and $c_2 = 1/(\nu-1)$. Therefore, in this case, $s^*(\omega) = (1-\omega)rc_1 + \omega c_2$.

Even the SSVS prior discussed in Section 2.3.1 can be written is this general form. From Equation (2.23), which uses the original notation, we see that $c_j^2\tau_j^2 = Q$ and $\tau_j^2 = rQ$, that is, adopting a Normal mixture for β_j, we have that $r = 1/c_j^2$ and the constant $c = 1$. The difference from the other mentioned priors is that SSVS prior does not uses an absolutely continuous prior for ψ_j as noted before.

Therefore, knowing the constant c for each prior (and assuming the ratio r and the other hyperparametes a and ν as fixed) allows us to generate $Q|\omega$ in a MCMC sampling scheme.

Table 2.1 gives a summary for what as discussed through Section 2.3 assuming the general form from (2.34) and viewing each prior as a scaled mixture of Normals (SMN). This completes our specification for the spike-and-slab mixture priors. The next section is dedicated to shrinking priors (and their frequentist analogues) that alone (without embedding in a mixture structure) can regularize the least square or maximum likelihood estimate, including the Normal-Gamma prior, which was mentioned above.

Prior	**Spike** $\psi\vert J=0$	**Slab** $\psi\vert J=1$	**Marginal** $\beta\vert\omega$	**Constant** c
SSVS	$\psi\vert J=0=\delta_{rQ}(.)$	$\psi\vert J=1=\delta_{Q}(.)$	$\omega\mathcal{N}(0,Q)+(1-\omega)\mathcal{N}(0,rQ)$	1
NMIG	$\mathcal{IG}(\nu,rQ)$	$\mathcal{IG}(\nu,Q)$	$\omega t_{2\nu}(0,Q/\nu)+(1-\omega)t_{2\nu}(0,rQ/\nu)$	$1/(\nu-1)$
Mixture of Laplaces	$\mathcal{E}(1/2rQ)$	$\mathcal{E}(1/2Q)$	$\omega Lap(\sqrt{Q})+(1-\omega)Lap(\sqrt{rQ})$	2
Mixture of Normal-Gammas	$\mathcal{G}(a,1/2rQ)$	$\mathcal{G}(a,1/2Q)$	$\omega\mathcal{NG}(\beta_j\vert a,Q)+(1-\omega)\mathcal{NG}(\beta_j\vert a,r,Q)$	$2a$
Laplace-t	$\mathcal{E}(1/2rQ)$	$\mathcal{IG}(\nu,Q)$	$\omega t_{2\nu}(0,Q/\nu)+(1-\omega)Lap(\sqrt{rQ})$	$c_1=2,\quad c_2=1/(\nu-1)$

Table 2.1: *Summary table: spike-and-slab mixture priors*

Chapter 3

Sparsity in dynamic linear models

In this chapter some basic concepts underlying the general dynamic linear model theory are introduced and developed in the context of the Gaussian *dynamic linear model* (DLM), which is presented as a special case of a general *state space model* (SSM), being linear and Gaussian.

In Section 3.1, we discuss the basic formulation of state space models and the structure of the recursive computations for estimation and prediction.

In Section 3.2, the specific case of the Gaussian DLM is presented (see e.g., Petris et al. (2009); West and Harrison (1997)) as well as the filtering and smoothing recursions for this case. In particular, we present the *Kalman filter*, the *Kalman smoother* and the *forward filtering backward sampling* (FFBS) algorithm for Gaussian DLMs.

Finally, we discuss some existing regularization methods for *time varying parameter* (TVP) models in Section 3.3. This completes the concepts needed to understand the model that will be presented in Chapter 5.

3.1 State space models

State space models (SSM) originated in the early sixties in the area of control engineering (Kalman et al. (1960)). It provides a general framework for analyzing deterministic and stochastic dynamical systems that are measured or observed through a stochastic process. The SSM framework has been successfully applied in engineering, statistics, computer science and economics to solve a broad range of dynamical systems problems. It appeared in the time series literature in the seventies (Akaike (1974); Harrison and Stevens (1976)) and became established during the eighties (Harvey (1989);West and Harrison (1997)).

The main applications in statistics are structural time series models and dynamic regression models of the form

$$y_t = \mu_t + \gamma_t + x_t\beta_t + \epsilon_t,$$

where μ_t is a trend component, γ_t is a seasonal component, β_t is the time-varying effect of the covariate x_t, and ϵ_t is an error. Gathering μ_t, γ_t and β_t into a vector called the states' vector and defining appropriate transition structures for them, these models can be written in a state space form (see, e.g., Harvey (1989)).

As noted in Petris et al. (2009), SSMs consider a time series as the output of a dynamic system perturbed by random disturbances. They allow a natural interpretation of a time series as the combination of several components, such as trend, seasonal or regressive components. The problems of estimation and forecasting are solved by recursively computing the conditional distribution of the hidden states, given the available information. In this sense, they are quite naturally treated within a Bayesian framework.

Consider a time series $\{y_t\}_{t\geq 1}$. Specifying the joint distribution of $(y_1, y_2, .., y_t)$ is not easy since in time series analysis the assumptions of independence are seldom justified. In that way, Markovian dependence is often assumed, i.e., we say that $\{y_t\}_{t\geq 1}$ is a Markov chain if, for any $t > 1$,

$$p(y_t|y_{1:t-1}) = p(y_t|y_{t-1}).$$

Another way of expressing the Markovian dependence is saying that y_t and $y_{1:t-2}$ are conditionally independent given y_{t-1}. Thus the joint distribution of $(y_1, y_2, .., y_t)$ is

$$p(y_{1:t}) = p(y_1)\prod_{t>1} p(y_t|y_{t-1}).$$

The following definition presents the assumptions that characterize a general state space model.

Definition 3.1. ***State space models.*** *Formally, a state space model (SSM) consists of an $\mathbb{R}^q$-valued unobserved time series $\{\boldsymbol{\theta}_t\}_{t\geq 0}$ (the states) and a $\mathbb{R}^m$-valued time series $\{\boldsymbol{y}_t\}_{t\geq 1}$ (the observations) satisfying the assumptions:*

(A.1) $\{\boldsymbol{\theta}_t\}_{t\geq 0}$ is a Markov chain;

(A.2) Conditionally on $\{\boldsymbol{\theta}_t\}_{t\geq 0}$, the $\{\boldsymbol{y}_t\}_{t\geq 1}$ are independent and $\boldsymbol{y}_t$ depends on $\boldsymbol{\theta}_t$ only.

State space models in which the states are discrete-valued random variables are often called hidden Markov models.

Throughout this chapter, we assume $m = 1$, i.e., we will work with univariate time series, where y_t is a scalar rather than a vector. Nevertheless, the derived results can easily be extended to multivariate time series.

From Definition 3.1, we conclude that a SSM is completely specified by the initial distribution of $p(\boldsymbol{\theta}_0)$ and the conditional densities $p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})$ and $p(y_t|\boldsymbol{\theta}_t)$. Thus, for $t > 0$, the joint distribution is as follows

$$p(\boldsymbol{\theta}_{0:t}, y_{1:t}) = p(\boldsymbol{\theta}_0)\prod_{t\geq 1} p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})p(y_t|\boldsymbol{\theta}_t). \tag{3.1}$$

It follows from the Definition 3.1 that $p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{0:t-1}, y_{1:t-1}) = p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})$.

The information flow assumed by a state space model is represented in Figure 3.1. The graph in the figure is a special case of a directed acyclic graph (DAG).

For a SSM, the main tasks are to make inference on the unobserved states or predict future observations based on a part of the observation sequence. To estimate de state vector one has

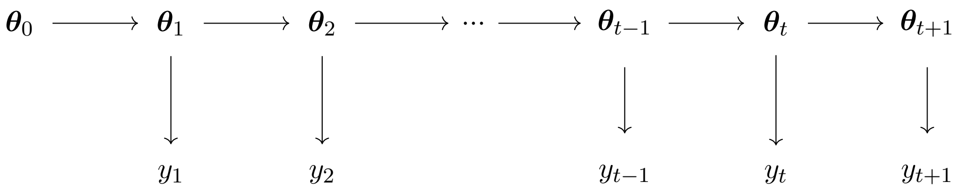


Figure 3.1: *Dependence structure for a state space model*

to compute the conditional densities $p(\boldsymbol{\theta}_s|y_{1:t})$ when $s = t$ (filtered density), $s < t$ (smoothed density) or $s > t$ (predictive density).

In the filtering problem, the data is supposed to arrive sequentially in time. In this case we want to update our current inference on the state vector as new data become available, that is, we want to estimate the filtering densities $p(\boldsymbol{\theta}_t|y_{1:t})$, then $p(\boldsymbol{\theta}_{t+1}|y_{1:t+1})$, and so on.

Conversely, the problem of smoothing, or retrospective analysis, consists in estimating the state sequence at times $1, ..., t$ given data $y_1, ..., y_t$, that is, estimate $p(\boldsymbol{\theta}_{1:t}|y_{1:t})$.

Finally, the problem of forecasting or prediction consists in estimating y_{t+h} based on data $y_{1:t}$, where h is the number of steps ahead. For example, for one-step-ahead forecasting, one has to estimate first the next value $\boldsymbol{\theta}_{t+1}$ of the state vector, and then, based on this estimate, compute the forecast for y_{t+1}. The one-step-ahead predictive density $p(\boldsymbol{\theta}_{t+1}|y_{1:t})$ is calculated through the filtering density $p(\boldsymbol{\theta}_t|y_{1:t})$. Then, one can calculate $p(y_{t+1}|y_{1:t})$.

One of the advantages of SSM is that, due to the Markovian structure of the state dynamics (A.1) and the assumption on the conditional independence for the observables (A.2), the filtered, smoothed and predictive densities can be computed using a recursive algorithm. The following proposition presents the filtering recursions for a general SSM.

Proposition 3.1. ***Filtering recursions.*** *For a general state space model defined by 3.1 the following statements hold. Starting from $\boldsymbol{\theta}_0 \sim p(\boldsymbol{\theta}_0)$, one can recursively compute, for $t \geq 1$:*

(i) The one-step ahead predictive density for the states $p(\boldsymbol{\theta}_t|y_{1:t-1})$ from the filtered density $p(\boldsymbol{\theta}_{t-1}|y_{1:t-1})$ as

$$p(\boldsymbol{\theta}_t|y_{1:t-1}) = \int p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})p(\boldsymbol{\theta}_{t-1}|y_{1:t-1})d\boldsymbol{\theta}_{t-1}.$$

(ii) The one-step ahead predictive density for the observations $p(y_t|y_{1:t-1})$ from the predictive density for the states $p(\boldsymbol{\theta}_t|y_{1:t-1})$ as

$$p(y_t|y_{1:t-1}) = \int p(y_t|\boldsymbol{\theta}_t)p(\boldsymbol{\theta}_t|y_{1:t-1})d\boldsymbol{\theta}_t.$$

(iii) The filtering density $p(\boldsymbol{\theta}_t|y_{1:t})$ from the above densities as

$$p(\boldsymbol{\theta}_t|y_{1:t}) = \frac{p(y_t|\boldsymbol{\theta}_t)p(\boldsymbol{\theta}_t|y_{1:t-1})}{p(y_t|y_{1:t-1})}.$$

Proof. Because $\boldsymbol{\theta}_t$ is conditionally independent of $y_{1:t-1}$ given $\boldsymbol{\theta}_{t-1}$, then

$$\begin{aligned}
p(\boldsymbol{\theta}_t|y_{1:t-1}) &= \int p(\boldsymbol{\theta}_{t-1}, \boldsymbol{\theta}_t|y_{1:t-1})d\boldsymbol{\theta}_{t-1} \\
&= \int p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}, y_{1:t-1})p(\boldsymbol{\theta}_{t-1}|y_{1:t-1})d\boldsymbol{\theta}_{t-1} \\
&= \int p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})p(\boldsymbol{\theta}t-1|y_{1:t-1})d\boldsymbol{\theta}_{t-1}.
\end{aligned}$$

Note also that y_t is conditionally independent of $y_{1:t-1}$ given $\boldsymbol{\theta}_t$. Therefore,

$$\begin{aligned}
p(y_t|y_{1:t-1}) &= \int p(y_t, \boldsymbol{\theta}_t|y_{1:t-1})d\boldsymbol{\theta}_t \\
&= \int p(y_t|\boldsymbol{\theta}_t, y_{1:t-1})p(\boldsymbol{\theta}_t|y_{1:t-1})d\boldsymbol{\theta}_t \\
&= \int p(y_t|\boldsymbol{\theta}_t)p(\boldsymbol{\theta}_t|y_{1:t-1})d\boldsymbol{\theta}_t.
\end{aligned}$$

Finally, to prove (iii) note that by Bayes' theorem

$$p(\boldsymbol{\theta}_t|y_{1:t}) = \frac{p(y_{1:t}, \boldsymbol{\theta}_t)p(\boldsymbol{\theta}_t)}{p(y_{1:t})} = \frac{p(y_t|\boldsymbol{\theta}_t, y_{1:t-1})p(y_{1:t-1}|\boldsymbol{\theta}_t)p(\boldsymbol{\theta}_t)}{p(y_t|y_{1:t-1})p(y_{1:t-1})} = \frac{p(y_t|\boldsymbol{\theta}_t, y_{1:t-1})p(\boldsymbol{\theta}_t|y_{1:t-1})p(y_{1:t-1})}{p(y_t|y_{1:t-1})p(y_{1:t-1})},$$

and by conditional independence of y_t and $y_{1:t-1}$ given $\boldsymbol{\theta}_t$, $p(y_t|\boldsymbol{\theta}_t, y_{1:t-1}) = p(y_t|\boldsymbol{\theta}_t)$. Then, we get the result (iii). Note that $p(\boldsymbol{\theta}_t|y_{1:t-1})$ is the prior distribution and $p(y_t|\boldsymbol{\theta}_t)$ is the likelihood. □

Using Proposition 3.1, k-steps ahead predictive distributions for the state and for the observation can be derived recursively according to

$$p(\boldsymbol{\theta}_{t+k}|y_{1:t}) = \int p(\boldsymbol{\theta}_{t+k}|\boldsymbol{\theta}_{t+k-1})p(\boldsymbol{\theta}_{t+k-1}|y_{1:t})d\boldsymbol{\theta}_{t+k-1}$$

and

$$p(y_{t+k}|y_{1:t}) = \int p(y_{t+k}|\boldsymbol{\theta}_{t+k})p(\boldsymbol{\theta}_{t+k}|y_{1:t})d\boldsymbol{\theta}_{t+k}.$$

Note that $p(\boldsymbol{\theta}_t|y_{1:t})$ summarizes the information contained in the past observations $y_{1:t}$, which is sufficient for predicting y_{t+k} for any $k > 0$.

Proposition 3.1 is about filtering and forecasting problems. In addition, if we want to reconstruct the retrospective behavior of the system given all available data up to a certain time T (the smoothing problem), we can use a backward-recursive algorithm to compute the conditional distributions of $\boldsymbol{\theta}_t$ given $y_{1:T}$ for any $t < T$, starting from the filtered density $p(\boldsymbol{\theta}_T|y_{1:T})$. The following proposition presents the smoothing recursion for a general SSM.

Proposition 3.2. ***Smoothing recursion.*** *For a general state space model defined by 3.1 the following statements hold.*

(i) Conditional on $y_{1:T}$ the state sequence $(\boldsymbol{\theta}_0, ..., \boldsymbol{\theta}_T)$ has the following backward transition probabilities for any $t < T$

$$p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t+1}, y_{1:T}) = \frac{p(\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t)p(\boldsymbol{\theta}_t|y_{1:t})}{p(\boldsymbol{\theta}_{t+1}|y_{1:t})}.$$

(ii) The smoothing distributions of $\boldsymbol{\theta}_t$ given $y_{1:T}$ can be computed by the following backward recursion in t, starting from $p(\boldsymbol{\theta}_T|y_{1:T})$

$$p(\boldsymbol{\theta}_t|y_{1:T}) = p(\boldsymbol{\theta}_t|y_{1:t}) \int \frac{p(\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t)}{p(\boldsymbol{\theta}_{t+1}|y_{1:t})} p(\boldsymbol{\theta}_{t+1}|y_{1:T}) d\boldsymbol{\theta}_{t+1}.$$

Proof. First, by the Markovian structure of a SSM, note that $\boldsymbol{\theta}_t$ and $y_{t+1:T}$ are conditionally independent given $\boldsymbol{\theta}_{t+1}$ and that $\boldsymbol{\theta}_{t+1}$ and $y_{1:T}$ are conditionally independent given $\boldsymbol{\theta}_t$. Them, using Bayes theorem we have

$$\begin{aligned} p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t+1}, y_{1:T}) &= p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t+1}, y_{1:t}) \\ &= \frac{p(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1}, y_{1:t})}{p(\boldsymbol{\theta}_{t+1}, y_{1:t})} = \frac{p(\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t, y_{1:t})p(\boldsymbol{\theta}_t|y_{1:t})}{p(\boldsymbol{\theta}_{t+1}|y_{1:t})} \\ &= \frac{p(\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t)p(\boldsymbol{\theta}_t|y_{1:t})}{p(\boldsymbol{\theta}_{t+1}|y_{1:t})}. \end{aligned}$$

To prove (ii) compute de marginal distribution $p(\boldsymbol{\theta}_t|y_{1:T})$ by

$$\begin{aligned} p(\boldsymbol{\theta}_t|y_{1:T}) &= \int p(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1}|y_{1:T}) d\boldsymbol{\theta}_{t+1} \\ &= \int p(\boldsymbol{\theta}_{t+1}|y_{1:T}) p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t+1}, y_{1:T}) d\boldsymbol{\theta}_{t+1} \\ &= \int p(\boldsymbol{\theta}_{t+1}|y_{1:T}) \frac{p(\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t)p(\boldsymbol{\theta}_t|y_{1:t})}{p(\boldsymbol{\theta}_{t+1}|y_{1:t})} d\boldsymbol{\theta}_{t+1} \\ &= p(\boldsymbol{\theta}_t|y_{1:t}) \int p(\boldsymbol{\theta}_{t+1}|y_{1:T}) \frac{p(\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t)}{p(\boldsymbol{\theta}_{t+1}|y_{1:t})} d\boldsymbol{\theta}_{t+1}, \end{aligned}$$

where we use (i) in the penultimate equality.

□

3.2 Dynamic linear models

3.2.1 Definition of dynamic linear models and the Kalman recursions

An important class of state space models is given by Gaussian linear state space models as stated in the following definition. We will se that the computation of conditional distributions simplify considerably under the Gaussian assumption.

Definition 3.2. ***Gaussian dynamic linear models.*** *The Gaussian dynamic linear model (DLM) is specified by a Normal prior distribution for the q-dimensional state vector at time $t = 0$,*

$$\boldsymbol{\theta}_0 \sim \mathcal{N}(\boldsymbol{m}_0, \boldsymbol{C}_0), \tag{3.2}$$

and the pair of equations for each time $t \geq 1$,

$$\boldsymbol{y}_t = \boldsymbol{F}_t' \boldsymbol{\theta}_t + \boldsymbol{\nu}_t, \qquad \boldsymbol{\nu}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{V}_t) \tag{3.3}$$

$$\boldsymbol{\theta}_t = \boldsymbol{G}_t \boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t, \qquad \boldsymbol{\omega}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{W}_t), \tag{3.4}$$

where $\boldsymbol{G}_t$ and $\boldsymbol{F}_t'$ are known matrices (of order $(q \times q)$ and $(m \times q)$ respectively) and $\{\boldsymbol{\nu}_t\}_{t\geq 1}$ and $\{\boldsymbol{\omega}_t\}_{t\geq 1}$ are two independent sequences of independent Gaussian vectors with mean zero and known variances $\{\boldsymbol{V}_t\}_{t\geq 1}$ and $\{\boldsymbol{W}_t\}_{t\geq 1}$.

Equation (3.3) is called the observation equation, while (3.4) is the evolution, state or system equation. Furthermore, it is assumed that $\boldsymbol{\theta}_0$ is independent of the errors $\boldsymbol{\nu}_t$ and $\boldsymbol{\omega}_t$ for any t.

Even though it is possible to assume heavy tailed errors for the general DLM, the price to be paid when removing the Normality assumption is additional computation difficulties. The assumption of Normality is sensible in many applications, though, it can be justified by central limit theorem arguments.

West and Harrison (1997) pointed out that more general models could be obtained by allowing the error sequences $\{\boldsymbol{\nu}_t\}$ and $\{\boldsymbol{\omega}_t\}$ to be both autocorrelated and cross correlated, and some definitions of dynamic linear models would allow for this structure. However, it is always possible to reformulate such a correlated model in terms of one that satisfies the independence assumptions. Thus, nothing is lost by imposing this restriction.

It is straightforward to show that the DLM of Definition 3.2 satisfies assumptions (A.1) and (A.2) that characterized a SSM in the Definition 3.1, with

$$\begin{aligned} \boldsymbol{y}_t|\boldsymbol{\theta}_t &\sim \mathcal{N}(\boldsymbol{F}_t'\boldsymbol{\theta}_t, \boldsymbol{V}_t), \\ \boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1} &\sim \mathcal{N}(\boldsymbol{G}_t\boldsymbol{\theta}_{t-1}, \boldsymbol{W}_t). \end{aligned} \tag{3.5}$$

In summary, the Gaussian DLM from Definition 3.2 is completely characterized by the set of quadruples $\{\boldsymbol{F}, \boldsymbol{G}, V, \boldsymbol{W}\}_t = \{\boldsymbol{F}_t, \boldsymbol{G}_t, \boldsymbol{V}_t, \boldsymbol{W}_t\}$ for each time t.

Of special interest are the following two subsets of the general class of DLMs: (i) If the pair $\{\boldsymbol{F}, \boldsymbol{G}\}_t$ is constant for all t then the model is referred to as a time series DLM, or TSDLM; (ii) A TSDLM whose observation variances $\boldsymbol{V}$ and evolution variances $\boldsymbol{W}$ are constant for all t is referred to as a constant DLM.

Again, from now on we assume $m = 1$, i.e., a univariate Gaussian DLM so that y_t, ν_t and V_t are scalars. Nevertheless, it is straightforward to derive the results that will be showed for multivariate time series.

As the Gaussian DLM is a special case of a SSM, the filtering and the forecasting problems can also be solved by the general recursions presented in the Proposition 3.1. In this case, it can be proved that the random vector $(\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_t, y_1, ..., y_t)$ has a Gaussian distribution for any $t \geq 1$. It follows that the marginal and conditional distributions are also Gaussian. Therefore, they are completely determined by their means and variances. The filtering recursions for Gaussian DLMs is given by the Kalman filter as follows.

Proposition 3.3. ***Kalman filter.*** *Consider the Gaussian DLM specified by 3.2. Let the posterior at time $t-1$ be $\boldsymbol{\theta}_{t-1}|y_{1:t-1} \sim \mathcal{N}(\boldsymbol{m}_{t-1}, \boldsymbol{C}_{t-1})$. Then, the following statements hold.*

(i) The one-step ahead predictive distribution of $\boldsymbol{\theta}_t$ given $y_{1:t-1}$ (i.e., the prior at t) is

$$\boldsymbol{\theta}_t|y_{1:t-1} \sim \mathcal{N}(\boldsymbol{a}_t, \boldsymbol{R}_t), \tag{3.6}$$

where

$$\boldsymbol{a}_t = \boldsymbol{G}_t \boldsymbol{m}_{t-1}, \qquad \boldsymbol{R}_t = \boldsymbol{G}_t \boldsymbol{C}_{t-1} \boldsymbol{G}_t' + \boldsymbol{W}_t.$$

(ii) The one-step ahead predictive distribution of y_t given $y_{1:t-1}$ is

$$y_t | y_{1:t-1} \sim \mathcal{N}(f_t, Q_t), \tag{3.7}$$

where

$$f_t = \boldsymbol{F}_t' \boldsymbol{a}_t, \qquad Q_t = \boldsymbol{F}_t' \boldsymbol{R}_t \boldsymbol{F}_t + V_t.$$

(iii) The filtering distribution of $\boldsymbol{\theta}_t$ given $y_{1:t}$ (i.e., the posterior at t) is

$$\boldsymbol{\theta}_t | y_{1:t} \sim \mathcal{N}(\boldsymbol{m}_t, \boldsymbol{C}_t), \tag{3.8}$$

where

$$\boldsymbol{m}_t = \boldsymbol{a}_t + \boldsymbol{R}_t \boldsymbol{F}_t Q_t^{-1} e_t, \qquad \boldsymbol{C}_t = \boldsymbol{R}_t - \boldsymbol{R}_t \boldsymbol{F}_t Q_t^{-1} \boldsymbol{F}_t' \boldsymbol{R}_t,$$

with $e_t = y_t - f_t$ (i.e., the forecast error).

Proof. The random vector $(\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_t, y_1, ..., y_t)$ has joint distribution given by Equation (3.1), where the marginal and conditional distributions involved are Gaussian. It follows that the joint distribution of $(\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_t, y_1, ..., y_t)$ is Gaussian for any $t \geq 1$. Therefore, from multivariate Normal standard results, the distribution of any subvector is Gaussian. Thus, the filtering and predictive distributions are Gaussian, so it sufficient to calculate its means and variances.

(i) Let $\boldsymbol{\theta}_t | y_{1:t-1} \sim \mathcal{N}(\boldsymbol{a}_t, \boldsymbol{R}_t)$. Using the system equation from Definition 3.2, we obtain

$$\begin{aligned} \boldsymbol{a}_t &= \mathbb{E}(\boldsymbol{\theta}_t | y_{1:t-1}) = \mathbb{E}(\mathbb{E}(\boldsymbol{\theta}_t | \boldsymbol{\theta}_{t-1}, y_{1:t-1}) | y_{1:t-1}) \\ &= \mathbb{E}(\boldsymbol{G}_t \boldsymbol{\theta}_{t-1} | y_{1:t-1}) = \boldsymbol{G}_t \boldsymbol{m}_{t-1}, \end{aligned}$$

and

$$\begin{aligned} \boldsymbol{R}_t &= \mathrm{Var}(\boldsymbol{\theta}_t | y_{1:t-1}) \\ &= \mathbb{E}(\mathrm{Var}(\boldsymbol{\theta}_t | \boldsymbol{\theta}_{t-1}, y_{1:t-1}) | y_{1:t-1}) + \mathrm{Var}(\mathbb{E}(\boldsymbol{\theta}_t | \boldsymbol{\theta}_{t-1}, y_{1:t-1}) | y_{1:t-1}) \\ &= \boldsymbol{W}_t + \boldsymbol{G}_t \boldsymbol{C}_{t-1} \boldsymbol{G}_t'. \end{aligned}$$

(ii) Let $y_t | y_{1:t-1} \sim \mathcal{N}(f_t, Q_t)$. Using the observation equation from Definition 3.2, we obtain

$$f_t = \mathbb{E}(y_t | y_{1:t-1}) = \mathbb{E}(\mathbb{E}(y_t | \boldsymbol{\theta}_t, y_{1:t-1}) | y_{1:t-1}) = \mathbb{E}(\boldsymbol{F}_t' \boldsymbol{\theta}_t | y_{1:t-1}) = \boldsymbol{F}_t' \boldsymbol{a}_t$$

and

$$\begin{aligned} Q_t &= \mathrm{Var}(y_t | y_{1:t-1}) \\ &= \mathbb{E}(\mathrm{Var}(y_t | \boldsymbol{\theta}_t, y_{1:t-1}) | y_{1:t-1}) + \mathrm{Var}(\mathbb{E}(y_t | \boldsymbol{\theta}_t, y_{1:t-1}) | y_{1:t-1}) \\ &= V_t + \boldsymbol{F}_t' \boldsymbol{R}_t \boldsymbol{F}_t. \end{aligned}$$

(iii) The problem is the same as the Bayesian inference problem for the Gaussian linear model (see Section 2.1.2)

$$y_t = \boldsymbol{F}_t'\boldsymbol{\theta}_t + \nu_t, \qquad \nu_t \sim \mathcal{N}(0, V_t),$$

with the regression vector parameter $\boldsymbol{\theta}_t | y_{1:t-1} \sim \mathcal{N}(\boldsymbol{a}_t, \boldsymbol{R}_t)$ (the prior) and V_t is known. By Proposition 3.1 (iii), we have that $p(\boldsymbol{\theta}_t|y_{1:t}) \propto p(y_t|\boldsymbol{\theta}_t)p(\boldsymbol{\theta}_t|y_{1:t-1})$, where $p(y_t|\boldsymbol{\theta}_t)$ is the likelihood. In the Gaussian DLM case, all the distributions are Gaussian, so the posterior

$$\boldsymbol{\theta}_t|y_{1:t} \sim \mathcal{N}(\boldsymbol{m}_t, \boldsymbol{C}_t).$$

Applying the results from Equation (2.5), we have that

$$\boldsymbol{C}_t = (\boldsymbol{R}_t^{-1} + \boldsymbol{F}_t V_t^{-1}\boldsymbol{F}_t')^{-1}$$

and

$$\boldsymbol{m}_t = \boldsymbol{C}_t(\boldsymbol{F}_t V_t^{-1} y_t + \boldsymbol{R}_t^{-1}\boldsymbol{a}_t).$$

Using the Woodbury matrix identity (Woodbury (1950)), we get

$$\boldsymbol{C}_t = \boldsymbol{R}_t - \boldsymbol{R}_t\boldsymbol{F}_t(V_t + \boldsymbol{F}_t'\boldsymbol{R}_t\boldsymbol{F}_t)^{-1}\boldsymbol{F}_t'\boldsymbol{R}_t = \boldsymbol{R}_t - \boldsymbol{R}_t\boldsymbol{F}_t Q_t^{-1}\boldsymbol{F}_t'\boldsymbol{R}_t,$$

where we use (ii) for the last equality. Finally, using the above identity and (ii), we have

$$\boldsymbol{m}_t = \boldsymbol{a}_t + \boldsymbol{R}_t\boldsymbol{F}_t Q_t^{-1}(y_t - f_t).$$

□

We have seen that the state space models' filtering recursions can be easily computed for a Gaussian DLM. Thus, the smoothing recursion presented in the Proposition 3.2 can also be stated more explicitly in terms of means and variances of the smoothing distributions. This is introduced by the following proposition.

Proposition 3.4. ***Kalman smoother.*** *Consider the Gaussian DLM specified by Definition 3.2. If* $\boldsymbol{\theta}_{t+1}|y_{1:T} \sim \mathcal{N}(\boldsymbol{s}_{t+1}, \boldsymbol{S}_{t+1})$, *then*

$$\boldsymbol{\theta}_t|y_{1:T} \sim \mathcal{N}(\boldsymbol{s}_t, \boldsymbol{S}_t),$$

where

$$\boldsymbol{s}_t = \boldsymbol{m}_t + \boldsymbol{C}_t\boldsymbol{G}_{t+1}'\boldsymbol{R}_{t+1}^{-1}(\boldsymbol{s}_{t+1} - \boldsymbol{a}_{t+1}),$$
$$\boldsymbol{S}_t = \boldsymbol{C}_t - \boldsymbol{C}_t\boldsymbol{G}_{t+1}'\boldsymbol{R}_{t+1}^{-1}(\boldsymbol{R}_{t+1} - \boldsymbol{S}_{t+1})\boldsymbol{R}_{t+1}^{-1}\boldsymbol{G}_{t+1}\boldsymbol{C}_t.$$

Proof. From the properties of the multivariate Gaussian distribution, we have that $\boldsymbol{\theta}_t|y_{1:T}$ is Gaussian. Therefore, we only have to calculate its mean and variance, i.e.,

$$\boldsymbol{s}_t = \mathbb{E}(\boldsymbol{\theta}_t|y_{1:T}) = \mathbb{E}(\mathbb{E}(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t+1}, y_{1:T})|y_{1:T})$$

and

$$\boldsymbol{S}_t = \mathrm{Var}(\boldsymbol{\theta}_t|y_{1:T}) = \mathrm{Var}(\mathbb{E}(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t+1}, y_{1:T})|y_{1:T}) + \mathbb{E}(\mathrm{Var}(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t+1}, y_{1:T})|y_{1:T}).$$

As shown before, by conditional independence, $p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t+1}, y_{1:T}) = p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t+1}, y_{1:t})$. We can compute this distribution by using the Bayes theorem. In the Proposition 3.3, we have noted the equivalence to the Bayesian inference problem for the Gaussian linear model. In this case we can do the same by noting that the state equation

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{G}_{t+1}\boldsymbol{\theta}_t + \boldsymbol{\omega}_{t+1}, \qquad \boldsymbol{\omega}_{t+1} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{W}_{t+1})$$

is equivalent to a regression problem, where the prior is $\boldsymbol{\theta}_t|y_{1:t} \sim \mathcal{N}(\boldsymbol{m}_t, \boldsymbol{C}_t)$, the likelihood is $p(\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t, y_{1:t}) = p(\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t) \sim \mathcal{N}(\boldsymbol{G}_{t+1}\boldsymbol{\theta}_t, \boldsymbol{W}_{t+1})$, where $\boldsymbol{W}_{t+1}$ is known. Therefore, applying the same results from the Equation 2.5 and the Woodbury matrix identity (Woodbury (1950))

$$\mathbb{E}(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t+1}, y_{1:t}) = \boldsymbol{m}_t + \boldsymbol{C}_t\boldsymbol{G}'_{t+1}\boldsymbol{R}^{-1}_{t+1}(\boldsymbol{\theta}_{t+1} - \boldsymbol{a}_{t+1}) = \boldsymbol{h}_t$$

and

$$\mathrm{Var}(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t+1}, y_{1:t}) = \boldsymbol{C}_t - \boldsymbol{C}_t\boldsymbol{G}'_{t+1}\boldsymbol{R}^{-1}_{t+1}\boldsymbol{G}_{t+1}\boldsymbol{C}_t = \boldsymbol{H}_t.$$

It follows that

$$\boldsymbol{s}_t = \mathbb{E}(\mathbb{E}(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t+1}, y_{1:t})|y_{1:T}) = \boldsymbol{m}_t + \boldsymbol{C}_t\boldsymbol{G}'_{t+1}\boldsymbol{R}^{-1}_{t+1}(\boldsymbol{s}_{t+1} - \boldsymbol{a}_{t+1}),$$

and

$$\begin{aligned}\boldsymbol{S}_t &= \mathrm{Var}(\mathbb{E}(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t+1}, y_{1:t})|y_{1:T}) + \mathbb{E}(\mathrm{Var}(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t+1}, y_{1:t})|y_{1:T})\\ &= \boldsymbol{C}_t\boldsymbol{G}'_{t+1}\boldsymbol{R}^{-1}_{t+1}\boldsymbol{S}_{t+1}\boldsymbol{R}^{-1}_{t+1}\boldsymbol{G}_{t+1}\boldsymbol{C}_t + \boldsymbol{C}_t - \boldsymbol{C}_t\boldsymbol{G}'_{t+1}\boldsymbol{R}^{-1}_{t+1}\boldsymbol{G}_{t+1}\boldsymbol{C}_t\\ &= \boldsymbol{C}_t - \boldsymbol{C}_t\boldsymbol{G}'_{t+1}\boldsymbol{R}^{-1}_{t+1}(\boldsymbol{R}_{t+1} - \boldsymbol{S}_{t+1})\boldsymbol{R}^{-1}_{t+1}\boldsymbol{G}_{t+1}\boldsymbol{C}_t.\end{aligned}$$

□

3.2.2 Dynamic linear models with unknown parameters

Until now, we have assumed that the system matrices $\boldsymbol{F}_t$, $\boldsymbol{G}_t$, $\boldsymbol{W}_t$ and the variance V_t of the univariate Gaussian DLM were known. In fact, these matrices are rarely completely known, so let the matrices and the variance depend on a vector of unknown parameters $\boldsymbol{\Phi}$. Usually, $\boldsymbol{\Phi}$ is constant over time, however, it is possible that the vector of unknown parameters be time-varying, i.e., $\boldsymbol{\Phi} = \boldsymbol{\Phi}_t$.

Dynamic regression. For example, an important case of the Gaussian DLM is the dynamic multiple regression (through the origin) that links the response y_t to q regressors $\boldsymbol{X}_t = (X_{1t}, ..., X_{qt})$ at time t

$$\begin{aligned}y_t &= \boldsymbol{X}_t\boldsymbol{\beta}_t + \nu_t, &\qquad \nu_t &\sim \mathcal{N}(0, \sigma^2_t),\\ \boldsymbol{\beta}_t &= \boldsymbol{\beta}_{t-1} + \boldsymbol{\omega}_t, &\qquad \boldsymbol{\omega}_t &\sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{W}_t),\end{aligned} \tag{3.9}$$

for $t = 1, ..., n$, where $\boldsymbol{X}_t$ is the $(1 \times q)$ vector of regressors, $\boldsymbol{\beta}_t' = (\beta_{1t}, ..., \beta_{qt})$ is the $(q \times 1)$ vector of coefficients and ν_t and $\boldsymbol{\omega}_t$ are two independent sequences of independent Gaussian errors with mean zero and variances σ_t^2 and $\boldsymbol{W}_t$, respectively. Thus, this is a Gaussian DLM defined by 3.2 with $\boldsymbol{F}_t' = \boldsymbol{X}_t$, $\boldsymbol{\theta}_t = \boldsymbol{\beta}_t$ (the states), $\boldsymbol{G}_t = \boldsymbol{G} = \boldsymbol{I}_q$ and $V_t = \sigma_t^2$. Note that setting $\boldsymbol{W}_t = \boldsymbol{0}$ for all t is equivalent to $\boldsymbol{\beta}_t = \boldsymbol{\beta}$, i.e., the static regression.

Some points should be highlighted in (3.9). First, note that it is assumed a time-varying observational variance σ_t^2, for which can can be assigned a stochastic volatility process. Nevertheless, we could also assume a constant variance $\sigma^2 = \sigma^2$. Second, even though it is assumed a time-varying evolution matrix $\boldsymbol{W}_t$ for the states (coefficients), often it is supposed that the coefficients are independent and evolve in time as a random walk with constant evolution matrix $\boldsymbol{W} = \text{diag}(\omega_1, ..., \omega_q)$.

In addition, often the system matrices $\boldsymbol{F}_t$ and $\boldsymbol{G}_t$ are completely known, such as in some structural models, but very rarely are the covariance matrices V_t and $\boldsymbol{W}_t$. A simple case is when V_t and $\boldsymbol{W}_t$ are known up to a common scale factor, that is, $V_t = \sigma^2 \tilde{V}_t$ and $\boldsymbol{W}_t = \sigma^2 \tilde{\boldsymbol{W}}_t$, where σ^2 is unknown and $\tilde{V}_t$ and $\tilde{\boldsymbol{W}}_t$ are known. An interesting case is when $\tilde{V}_t = 1$ (considering univariate DLMs) or $\tilde{V}_t = I_m$ (for multivariate DLMs, where $\boldsymbol{y}_t \in \mathbb{R}^m$), and $\tilde{\boldsymbol{W}}_t$ is specified by a discount factor. For a good discussion on discount factors, see Chapter 4 from Petris et al. (2009) or Chapter 6 from West and Harrison (1997).

Bayesian approach for Gaussian DLM with unknown parameters $\boldsymbol{\Phi}$ assume prior knowledge about $\boldsymbol{\Phi}$, which is expressed through a prior distribution. It is also assumed that state space hypothesis for the processes y_t and $\boldsymbol{\theta}_t$ hold conditionally on the parameters $\boldsymbol{\Phi}$. Thus, the joint distribution for $t \geq 1$ is

$$(\boldsymbol{\theta}_{0:t}, y_{1:t}, \boldsymbol{\Phi}) \sim p(\boldsymbol{\theta}_0|\boldsymbol{\Phi})p(\boldsymbol{\Phi}) \prod_{t \geq 1} p(y_t|\boldsymbol{\theta}_t, \boldsymbol{\Phi})p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}, \boldsymbol{\Phi}), \tag{3.10}$$

which is similar to the previous Equation (3.1), except from the fact that now we have a prior distribution for the unknown parameters $\boldsymbol{\Phi}$ and all the probabilities are conditional on them. The filtering density is then given by

$$p(\boldsymbol{\theta}_t|y_{1:t}) = \int p(\boldsymbol{\theta}_t|\boldsymbol{\Phi}, y_{1:t})p(\boldsymbol{\Phi}|y_{1:t})d\boldsymbol{\Phi}.$$

Inference on the vector $\boldsymbol{\theta}_{0:t}$ and $\boldsymbol{\Phi}$ is expressed through their joint posterior distribution

$$p(\boldsymbol{\theta}_{0:t}, \boldsymbol{\Phi}|y_{1:t}) = p(\boldsymbol{\theta}_{0:t}|\boldsymbol{\Phi}, y_{1:t})p(\boldsymbol{\Phi}|y_{1:t}), \tag{3.11}$$

where, using conjugate priors, it can be computed in closed form. However, in most cases, the joint posterior is analytically intractable, so we often use *Markov chain Monte Carlo* (MCMC) methods (see, e.g., Gamerman and Lopes (2006)) or sequential Monte Carlo methods.

The inclusion of the states in the posterior distribution usually simplifies the sampling, even when one is only interested in the posterior distribution of the $\boldsymbol{\Phi}$. Indeed, drawing a random variable/vector from $p(\boldsymbol{\Phi}|\boldsymbol{\theta}_{0:t}, y_{1:t})$ is almost invariably much easier than drawing it from $p(\boldsymbol{\Phi}|y_{1:t})$. This suggests that a sample from the joint distribution (3.11) can be obtained by alternating draws from $p(\boldsymbol{\Phi}|\boldsymbol{\theta}_{0:t}, y_{1:t})$ and $p(\boldsymbol{\theta}_{0:t}|\boldsymbol{\Phi}, y_{1:t})$ in a Gibbs sampler.

Forward filtering backward sampling. While $p(\mathbf{\Phi}|\boldsymbol{\theta}_{0:t}, y_{1:t})$ is problem specific, $p(\boldsymbol{\theta}_{0:t}|\mathbf{\Phi}, y_{1:t})$ has a general expression. We have seen that the smoothing recursions from the Proposition 3.4 provide an algorithm for computing the mean and variance of the the distribution $\boldsymbol{\theta}_t|y_{1:T}$ considering the Gaussian DLM. Note that, using the chain rule from probability theory, we can write the joint distribution of $\boldsymbol{\theta}_{0:T}$ given $y_{1:T}$ as

$$p(\boldsymbol{\theta}_{0:T}|y_{1:T}) = \prod_{t=0}^{T} p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t+1:T}, y_{1:T}),$$

where the last factor is $p(\boldsymbol{\theta}_T|y_{1:T})$, which is the filtering distribution of $\boldsymbol{\theta}_T$. By Proposition 3.3, we know that $\boldsymbol{\theta}_T|y_{1:T} \sim \mathcal{N}(\boldsymbol{m}_T, \boldsymbol{C}_T)$. Therefore, one can start by drawing $\boldsymbol{\theta}_T$ using the Kalman filter and then, for $t = T-1, T-2, ..., 0$, recursively draw $\boldsymbol{\theta}_t$ using the smoothing recursions. From the proof of Proposition 3.4 we have seen that $p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t+1:T}, y_{1:T}) = p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t+1}, y_{1:t})$, where it was shown that this distribution is $\mathcal{N}(\boldsymbol{h}_t, \boldsymbol{H}_t)$, with

$$\begin{aligned}\boldsymbol{h}_t &= \boldsymbol{m}_t + \boldsymbol{C}_t\boldsymbol{G}'_{t+1}\boldsymbol{R}^{-1}_{t+1}(\boldsymbol{\theta}_{t+1} - \boldsymbol{a}_{t+1}),\\ \boldsymbol{H}_t &= \boldsymbol{C}_t - \boldsymbol{C}_t\boldsymbol{G}'_{t+1}\boldsymbol{R}^{-1}_{t+1}\boldsymbol{G}_{t+1}\boldsymbol{C}_t,\end{aligned}$$

where we note that $\boldsymbol{h}_t$ explicitly depends on the value of $\boldsymbol{\theta}_{t+1}$ generated previously.

This sampling method is widely known as forward filtering backward sampling (FFBS) and is due to Carter and Kohn (1994) and Frühwirth-Schnatter (1994). The method is summarized in the Algorithm 4.

Algorithm 4: Forward filtering backward sampling

```
1. Run one step of the Kalman filter.
2. Draw $\boldsymbol{\theta}_T|y_{1:T} \sim \mathcal{N}(\boldsymbol{m}_T, \boldsymbol{C}_T)$.
3. For $t = T-1, ..., 0$, draw $(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t+1:T}, y_{1:T}) \sim \mathcal{N}(\boldsymbol{h}_t, \boldsymbol{H}_t)$.
```

In the context of Gaussian DLMs with unknown parameters, the FFBS algorithm can be applied to sample from the joint posterior distribution presented in Equation (3.11). That is, it is enough to draw $\boldsymbol{\theta}^{(i)}_{0:T}$ from $p(\boldsymbol{\theta}_{0:T}|\mathbf{\Phi} = \mathbf{\Phi}^{(i-1)}, y_{1:T})$ using FFBS and then draw $\mathbf{\Phi}^{(i)}$ from $p(\mathbf{\Phi}|\boldsymbol{\theta}_{0:t} = \boldsymbol{\theta}^{(i)}_{0:T}, y_{1:t})$ to obtain the sample $\{(\boldsymbol{\theta}^{(i)}_{0:T}, \mathbf{\Phi}^{(i)}), i = 1, ..., M\}$, where M is the total number of iterations. The sampling strategy is summarized as follows.

Algorithm 5: Forward filtering backward sampling in a Gibbs sampler

```
1. Initialize $\mathbf{\Phi} = \mathbf{\Phi}^{(0)}$.
2. For $i = 1, ..., M$:
a) Draw $\boldsymbol{\theta}^{(i)}_{0:T}$ from $p(\boldsymbol{\theta}_{0:T}|\mathbf{\Phi} = \mathbf{\Phi}^{(i-1)}, y_{1:T})$ using FFBS;
b) Draw $\mathbf{\Phi}^{(i)}$ from $p(\mathbf{\Phi}|\boldsymbol{\theta}_{0:t} = \boldsymbol{\theta}^{(i)}_{0:T}, y_{1:t})$.
```

The above MCMC strategy will be used to draw the posterior distribution of the dynamic linear regression model regularized by a dynamic version of the NMIG prior. The hierarchical specification and the results will be presented in Chapter 5. In the next session, some existing sparsity inducing methods for DLMs are discussed in order to introduce the subject and prepare for the model proposed later.

3.3 Sparsity in time-varying parameter models

Chapter 2 was dedicated to shrinkage and sparsity inducing priors independently assigned to static coefficients from the Gaussian linear model. Now we turn attention to the case where the coefficients from regression are time-varying, the so-called time-varying parameter (TVP) regression models. There are many recent papers that have applied shrinking or parsimony-inducing methods for TVP models such as Frühwirth-Schnatter and Wagner (2010), Chan et al. (2012), Belmonte et al. (2014), Kalli and Griffin (2014), Lopes et al. (2014), Bitto and Frühwirth-Schnatter (2016) and a few others. In the following sections we discuss the proposed models of each of these references, dividing them by related topics.

3.3.1 Parsimony-inducing priors for state space models

Even though the terminology of TVP models is mainly related to the econometric literature of univariate and multivariate (vector autoregression or VAR) regression models, in this subsection we will refer to generic state space models. The method proposed by Frühwirth-Schnatter and Wagner (2010) is based on extending the Bayesian variable selection approach, which is usually applied to regression models, to state space models. The approach determines which components to include in the model and specifies whether these components are fixed or time-varying. For this reason it was called *stochastic model specification search* for Gaussian and partial non-Gaussian state space models.

The basic state space model considered in the paper is the dynamic linear trend model defined for $t = 1, ..., T$ as

$$y_t = \mu_t + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0, \sigma^2), \tag{3.12}$$

where the trend μ_t follows a random walk with a random drift a_t starting from unknown initial values μ_0 and a_0

$$\begin{aligned} \mu_t &= \mu_{t-1} + a_{t-1} + \omega_{1t}, \qquad & \omega_{1t} \sim \mathcal{N}(0, \theta_1), \\ a_t &= a_{t-1} + \omega_{2t}, \qquad & \omega_{2t} \sim \mathcal{N}(0, \theta_2). \end{aligned} \tag{3.13}$$

The model defined by (3.12) and (3.13) can be rewritten as follows. Define two independent random walk processes $\tilde{\mu}_t$ and $\tilde{a}_t$ with Normal standard independent increments as well as an integrated process $\tilde{A}_t$ as

$$\begin{aligned} \tilde{\mu}_t &= \tilde{\mu}_{t-1} + \tilde{\omega}_{1t}, \qquad & \tilde{\omega}_{1t} \sim \mathcal{N}(0, 1) \\ \tilde{a}_t &= \tilde{a}_{t-1} + +\tilde{\omega}_{2t}, \qquad & \tilde{\omega}_{2t} \sim \mathcal{N}(0, 1) \\ \tilde{A}_t &= \tilde{A}_{t-1} + \tilde{a}_{t-1}, & \end{aligned} \tag{3.14}$$

with $\tilde{\mu}_0 = \tilde{a}_0 = \tilde{A}_0 = 0$. Thus, combining (3.14) with the following observation equation

$$y_t = \mu_0 + ta_0 + \sqrt{\theta_1}\tilde{\mu}_t + \sqrt{\theta_2}\tilde{A}_t + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0, \sigma^2) \tag{3.15}$$

we get what the authors named *non-centered parametrization* of the dynamic linear trend model.

In order to verify this result, define

$$\begin{aligned}a_t &= a_0 + \sqrt{\theta_2}\tilde{a}_t,\\ \mu_t &= \mu_0 + ta_0 + \sqrt{\theta_1}\tilde{\mu}_t + \sqrt{\theta_2}\tilde{A}_t\end{aligned}$$

and calculate $a_t - a_{t-1}$ and $\mu_t - \mu_{t-1}$, for which we obtain the centered parametrization presented in the Equation (3.13).

The state space representation of the non-centered model defined by (3.14) and (3.15) is then

$$\begin{aligned}y_t &= \boldsymbol{z}_t'\boldsymbol{\alpha} + \boldsymbol{H}\boldsymbol{x}_t + \varepsilon_t, & \varepsilon_t &\sim \mathcal{N}(0,\sigma^2)\\ \boldsymbol{x}_t &= \boldsymbol{F}\boldsymbol{x}_{t-1} + \boldsymbol{w}_t, & \boldsymbol{w}_t &\sim \mathcal{N}(0,\boldsymbol{W}),\end{aligned} \tag{3.16}$$

where $\boldsymbol{x}_0 = \boldsymbol{0}$ and

$$\boldsymbol{x}_t = \begin{pmatrix}\tilde{\mu}_t\\ \tilde{a}_t\\ \tilde{A}_t\end{pmatrix}, \boldsymbol{F} = \begin{pmatrix}1&0&0\\0&1&0\\0&1&1\end{pmatrix}, \boldsymbol{W} = \begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix},$$

$$\boldsymbol{H} = \begin{pmatrix}\sqrt{\theta_1} & 0 & \sqrt{\theta_2}\end{pmatrix}, \qquad \boldsymbol{z}_t = \begin{pmatrix}1 & t\end{pmatrix}, \qquad \boldsymbol{\alpha} = \begin{pmatrix}\mu_0 & a_0\end{pmatrix}.$$

Theoretically, this state space form could be used to perform Kalman filter (Proposition 3.3) and to compute the integrated likelihood

$$p(y_{1:T}|\boldsymbol{\Phi}) = \int p(y_{1:T}|\boldsymbol{x}_{1:T},\boldsymbol{\Phi})d\boldsymbol{x}_{1:T},$$

where $\boldsymbol{\Phi} = (\sqrt{\theta_1}, \sqrt{\theta_2}, \sigma^2, \mu_0, a_0)$.

However, the non-centered parametrization of dynamic linear trend model is not identified because in the observation equation (3.15) the sign of $\sqrt{\theta_1}$ and the sequence $\tilde{\mu}_{1:T}$ may be changed by multiplying all elements with -1 without changing the distribution of $y_{1:T}$. That is, if we define the state $\boldsymbol{x}_t^* = (-\tilde{\mu}_t, \tilde{a}_t, \tilde{A}_t)'$ and $\boldsymbol{\Phi}^* = (-\sqrt{\theta_1}, \sqrt{\theta_2}, \sigma^2, \mu_0, a_0)$ we have that

$$p(y_{1:T}|\boldsymbol{\Phi}) = \int p(y_{1:T}|\boldsymbol{x}_{1:T},\boldsymbol{\Phi})d\boldsymbol{x}_{1:T} = \int p(y_{1:T}|\boldsymbol{x}^*_{1:T},\boldsymbol{\Phi}^*)d\boldsymbol{x}^*_{1:T} = p(y_{1:T}|\boldsymbol{\Phi}^*).$$

Similarly, the sign of $\sqrt{\theta_2}$ and the sequences $\tilde{a}_t$ and $\tilde{A}_t$ may be changed without changing the integrated likelihood.

As a consequence, the function $p(y_{1:T}|\boldsymbol{\Phi})$ is symmetric around 0 in the direction of $\sqrt{\theta_1}$ and $\sqrt{\theta_2}$ and therefore multimodal. If the true variances θ_1 and θ_2 are positive, then the likelihood function concentrates around four modes. If one of these true variances is positive and the other is equal to 0, two of these modes collapse and the likelihood becomes bimodal with increasing T. If both true variances are equal to 0, then the likelihood function becomes unimodal as T increases. Thus, considering the non-centered parametrization and allowing for non-identifiability one can test whether the variances of the state space model are zero.

The multimodal property of the integrated likelihood gave an insight for building what the authors called the *parsimonious dynamic linear trend model.* The non-centered parametrization of this model is defined by the observation equation

$$y_t = \mu_0 + \delta t a_0 + \gamma_1\sqrt{\theta_1}\tilde{\mu}_t + \gamma_2\sqrt{\theta_2}\tilde{A}_t + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0,\sigma^2),$$

together with the Equation (3.14), where $(\delta, \gamma_1, \gamma_2)$ are the model binary indicators. While δ is responsible for including or not including trend (if $\delta = 0$, then the initial slope $a_0 = 0$), γ_1 and γ_2 decide if a certain component of the state vector is fixed (the case when $\sqrt{\theta}_i = 0$, $i = 1, 2$) or time-varying.

Evidently, $(\delta, \gamma_1, \gamma_2) = (1, 1, 1)$ corresponds to the unrestricted dynamic linear trend model defined by (3.12) and (3.13). The combination $(\delta, \gamma_1, \gamma_2) = (0, 1, 0)$ leads to the local linear model or exponential smoothing, the combination $(\delta, \gamma_1, \gamma_2) = (1, 0, 0)$ leads to a regression model with deterministic linear trend and $(\delta, \gamma_1, \gamma_2) = (0, 0, 0)$ leads to a i.i.d Normal data $y_t \sim \mathcal{N}(\mu_0, \sigma^2)$.

The state space representation of the non-centered parametrization of the parsimonious dynamic linear trend model is given by

$$\begin{aligned} y_t &= \boldsymbol{z}_t(\delta)'\boldsymbol{\alpha} + \boldsymbol{H}(\gamma_1, \gamma_2)\boldsymbol{x}_t + \varepsilon_t, & \varepsilon_t &\sim \mathcal{N}(0, \sigma^2) \\ \boldsymbol{x}_t &= \boldsymbol{F}\boldsymbol{x}_{t-1} + \boldsymbol{w}_t, & \boldsymbol{w}_t &\sim \mathcal{N}(0, \boldsymbol{W}), \end{aligned}$$

where $\boldsymbol{x}_t$, $\boldsymbol{F}$ and $\boldsymbol{\alpha}$ are the same as (3.16), while $\boldsymbol{H}(\gamma_1, \gamma_2)$ and $\boldsymbol{z}_t(\delta)$ depend on the model indicators

$$\boldsymbol{H}(\gamma_1, \gamma_2) = (\gamma_1\sqrt{\theta_1} \quad 0 \quad \gamma_2\sqrt{\theta_2}), \qquad \boldsymbol{z}_t(\delta) = (1 \quad \delta t).$$

Note that the parsimonious dynamic linear trend model has more unknown parameters than the unrestricted case because of the model binary indicators. That is, the vector of unknown parameters is now $\boldsymbol{\Phi} = (\sqrt{\theta}_1, \sqrt{\theta}_2, \sigma^2, \mu_0, a_0, \delta, \gamma_1, \gamma_2)$.

The prior distributions assigned for each parameter of $\boldsymbol{\Phi}$ is as follows. For the model indicators it is assumed an Uniform distribution over all possible 8 combinations, i.e., each combination is equally likely to be observed.

As common for dynamic linear trend models, it is assumed that a priori μ_0 and a_0 are independently normally distributed

$$\mu_0 \sim \mathcal{N}(y_1, P_{0,11}\sigma^2), \qquad a_0 \sim \mathcal{N}(0, P_{0,22}\sigma^2).$$

Furthermore, for the observation variance σ^2 it is assumed an inverted Gamma prior

$$\mathcal{IG}(c_0, C_0).$$

Lastly, in contrast with previous work, they did not use the usual Inverse-Gamma priors for the variances θ_1 and θ_2, i.e., $\theta_1 \sim \mathcal{IG}(d_{0,1}, D_{0,1})$ and $\theta_2 \sim \mathcal{IG}(d_{0,2}, D_{0,2})$, assuming Gaussian distributions for the standard deviations instead. That is, it was assumed a priori that

$$\sqrt{\theta}_1 \sim \mathcal{N}(0, B_{0,1}\sigma^2), \qquad \sqrt{\theta}_2 \sim \mathcal{N}(0, B_{0,2}\sigma^2).$$

Accordingly to the authors, the reason for choosing the Normal prior is based on the strong influence of the hyperparameters of the Inverse-Gamma prior on the posterior density of θ_i, if the true value θ_i^{tr} is close to 0. Due to the symmetry of the likelihood discussed above, the posterior density of $\pm\sqrt{\theta_i}$ (which is obtained from θ_i by multiplying the square root of θ_i with a random sign) is symmetric around zero as long as the prior is also symmetric around zero. If the unknown true variance θ_i^{tr} is significantly different from zero, then the posterior density of

$\pm\sqrt{\theta_i}$ is likely to be bimodal with the modes being close to $\pm\sqrt{\theta_i^{tr}}$. Otherwise, if the true θ_i^{tr} is close to or equal to zero, then the posterior density of $\pm\sqrt{\theta_i}$ is likely to be centered around zero.

Whereas the posterior is robust to the choice of hyperparameter $B_{0,i}$ in the Normal prior, it turns out to be rather sensitive to hyperparameter $D_{0,i}$ in the Inverse-Gamma prior for $i = 1, 2$. A practical example was given by fixing two values for $\theta_1 \in \{0.01, 0\}$ and assuming the priors $\theta_1 \sim \mathcal{IG}(0.5, D_{0,1})$ or $\sqrt{\theta_1} \sim \mathcal{N}(0, B_{0,1}\sigma^2)$ for various scale parameters $D_{0,1} \in \{0.015, 0.1, 0.2275\}$ and $B_{0,1} \in \{1, 10, 100\}$. When $\theta_1^{tr} = 0.01$ both posteriors are roughly the same and clearly indicate that $\theta_1^{tr} > 0$. A remarkable difference occurred when $\theta_1^{tr} = 0$. Under the Normal prior, the posterior of $\pm\sqrt{\theta_1}$ is centered at 0 for all values of $B_{0,1}$. However, under the Inverse-Gamma prior, the posterior is always centered away from 0, for all the values $D_{0,1}$.

Posterior inference is obtained by an MCMC approach, where the indicators $(\delta, \gamma_1, \gamma_2)$ and the parameters $(\sqrt{\theta_1}, \sqrt{\theta_2}, \sigma^2, \mu_0, a_0)$ are sampled jointly in one block and the states $\boldsymbol{x}_{1:T}$ are sampled using the FFBS (Algorithm 4). For details on the MCMC scheme, please refer to Section 2.4 from Frühwirth-Schnatter and Wagner (2010).

Another important reference on parsimony-inducing priors is Lopes et al. (2014), who proposed a mixture prior for the autoregressive parameters of the state equation. In the article, two specifications for the observation equation were considered: (i) a dynamic linear regression, and (ii) a standard stochastic volatility model. The mixture prior is presented as follows.

Consider the dynamic linear regression model from equation (3.9) with an intercept in the state equation. Assume that we have only one regressor for simplicity. Model (i) is then specified as

$$\begin{aligned} y_t &= x_t s_t + \eta_t \\ s_t &= \alpha + \beta s_{t-1} + \tau\varepsilon_t, \end{aligned}$$

where s_t is latent (hidden) state-space variable, η_t and ε_t are independent random shocks in the observation and state equations respectively, usually Gaussian, and the pair (x_t, y_t) is observed. The proposed mixture prior for (α, β, τ) is

$$\begin{aligned} p(\alpha, \beta, \tau) = &\, p_{01} p(\tau|\beta = 1)\delta_{\{\alpha=0,\beta=1\}} \\ &+ p_{00} p(\tau|\beta = 0)\delta_{\{\alpha=0,\beta=0\}} \\ &+ p_{u0} p(\tau|\beta = 0) p(\alpha|\beta = 0, \tau)\delta_{\{\beta=0\}} \\ &+ p_{uu} p(\beta) p(\tau|\beta \neq 0) p(\alpha|\beta), \end{aligned}$$

where p_{01}, p_{00}, p_{u0} and p_{uu} are the mixture weights of our four components. The notation δ_x represents the distribution such that x happens for sure, p_{01} is the probability that $(\alpha, \beta) = (0, 1)$, p_{00} is the probability that $(\alpha, \beta) = (0, 0)$, p_{u0} is the probability that $\beta = 0$ and α is unrestricted, and p_{uu} is the probability that $\beta \in (0, 1)$ and α is unrestricted. $p(\tau|\beta)$ denotes a discrete distribution on a grid which will be discussed below.

The prior was structured in order to express a small τ, which expresses the notion that the state evolves smoothly, while (α, β) can assume different values. In the paper, only four cases were considered. The cases as well as their interpretation are

- $(\alpha, \beta) = (0, 1)$, i.e., the state evolves like a random walk;
- $(\alpha, \beta) = (0, 0)$, i.e., the state is fixed near zero;

- $(\alpha, \beta) = (\alpha, 0)$, i.e., the state simply varies around a fixed level α, which is unrestricted;
- $(\alpha, \beta) = (\alpha, \beta)$, where $0 < \beta < 1$, i.e., the state varies in a stationary fashion;

It is important to note that zero prior weight was given on $\beta < 0$ and that the case $\tau = 0$ was not considered.

The priors for α, β, τ, s_0 was specified as follows. For τ, they chose a discrete prior over a n-dimensional grid of evenly spaced values $(t_1, t_2, .., t_n)$, with $t_1 = \tau_{min}$ (minimum value) and $t_n = \tau_{max}$ (maximum value). It was supposed that $p(\tau = \tau_{min}) = p_{min}$ and that, for $i > 1$, $p(\tau = t_i) \propto \exp(-c_\tau |t_i - \tau_{min}|)$. Thus, the prior for τ has four hyperparameters $(\tau_{min}, \tau_{max}, p_{min}, c_\tau)$. They allowed for the possibility that the choice of the four hyperparameters could depend on β. For example, one may want a smaller values of τ when $\beta = 0$. For this reason, the choice of c_τ given $\beta = 0$ was twice the value used for non-zero β.

The prior for α also depends on the value of β and τ, that is, it was assumed that $(\alpha|\beta, \tau) \sim \mathcal{N}(0, \sigma^2_\alpha(1 - \beta^2))$. As β increases, α is shrunken towards the case where $\alpha = 0$ at $\beta = 1$. When $\beta = 0$, it is simply assumed that $\alpha \sim \mathcal{N}(0, \sigma^2_\alpha)$.

To specify a prior for $\beta \in (0, 1)$, they used a grid of points $(b_1, b_2, ..., b_n)$, where they let $p(\beta = b_i) \propto \mathcal{N}(|\tilde{\beta}, \sigma^2_\beta)$.

Finally, for the initial state s_0, a mixture prior was assigned in order to induce sparsity since the value $s_0 = 0$ represents a model simplification in many cases. In fact, they used the SSVS prior presented in Section 2.3.1.

Posterior sitribution was calculated using an augmented Gibbs sampler. For details on the sampling scheme, please refer to Section 2.2 of Lopes et al. (2014).

3.3.2 Shrinking priors under the non-centered parametrization

The last section was dedicated to introduce some important references on parsimony-inducing priors which automatically choose some structural components of a state space model. In that way, those strategies are similar to the variable selection problem discussed in Section 2.3. In this section we discuss another strategy for inducing sparsity in time-varying parameter models that is more related to the regularization methods presented in Section 2.2. In particular, we discuss two alike papers which used the same kind of parametrization for the state equation.

In Belmonte et al. (2014) a TVP version of the generalized Phillips curve was estimated with the coefficients shrunken by a lasso type prior. Their basic model is formalized as a state space model

$$
\begin{aligned}
\pi_{t+h} &= \boldsymbol{\theta}^*_t \boldsymbol{z}_t + \varepsilon_{t+h} \\
\boldsymbol{\theta}^*_t &= \boldsymbol{\theta}^*_{t-1} + \boldsymbol{\eta}_t,
\end{aligned}
$$

where the variable of interest is the h-step-ahead inflation, defined as $\pi_{t+h} = \log(P_{t+h}) - \log(P_t)$, $\boldsymbol{z}_t = [1, \Delta \log(P_t), ..., \Delta \log(P_{t-p+1}), \boldsymbol{x}_t]$, $\boldsymbol{x}_t$ is a $(q \times 1)$ vector of exogenous predictors, and $\boldsymbol{\theta}^*_t = (\alpha'_t, \phi'_{t,0}, ..., \phi'_{t,p}, \gamma'_t)'$. It is assumed that $\varepsilon_t \sim \mathcal{N}(0, \sigma^2_t)$ and $\boldsymbol{\eta}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega})$, where $\boldsymbol{\Omega} = \text{diag}(\omega^2_1, ..., \omega^2_k)$, with $k = 1 + p + q$. The errors are supposed to be independent of each other and independent at all leads and lags. Note that it is allowed for heteroskedasticity in the observation equation, in particular, it is assumed a standard stochastic volatility specification for σ^2_t.

The initial condition for the states plays the role of a regression effect. Thus, the model can be rewritten breaking the coefficients into a constant part $\boldsymbol{\theta} = \boldsymbol{\theta}_0^*$ and a time-varying part $\boldsymbol{\theta}_t = \boldsymbol{\theta}_t^* - \boldsymbol{\theta}$ as follows

$$\begin{aligned} \pi_{t+h} &= \boldsymbol{\theta} \boldsymbol{z}_t + \boldsymbol{\theta}_t \boldsymbol{z}_t + \varepsilon_{t+h} \\ \boldsymbol{\theta}_t &= \boldsymbol{\theta}_{t-1} + \boldsymbol{\eta}_t \\ \boldsymbol{\theta}_0 &= \mathbf{0}. \end{aligned}$$

In order to induce shrinkage, the authors used another transformation: each state $\theta_{j,t}$ was divided by its standard deviation ω_j for $j = 1, .., k$, what was called non-centered parametrization and it was based on Frühwirth-Schnatter and Wagner (2010)

$$\begin{aligned} \pi_{t+h} &= \sum_{j=1}^{k} \theta_i z_{j,t} + \sum_{j=1}^{k} \omega_i \tilde{\theta}_{j,t} z_{j,t} + \varepsilon_{t+h} \\ \tilde{\theta}_{j,t} &= \tilde{\theta}_{j,t-1} + u_{j,t} \\ \tilde{\theta}_{j,0} &= 0, \end{aligned} \tag{3.17}$$

where $\tilde{\theta}_{j,t} = \theta_{j,t}/\omega_j$ and $u_{j,t} \sim \mathcal{N}(0,1)$ for $j = 1, ..., k$. Note that ω_j is not time-varying.

The prior used to shrink the coefficients is defined as follows. For the constant coefficients $\boldsymbol{\theta} = (\theta_1, ..., \theta_k)'$, it is assigned an hierarchical mixture of normal priors with an exponential mixing distribution inspired by the traditional Bayesian lasso (see Equation (2.17)). In particular, each θ_j for $j = 1, ..., k$ is assumed to be, a priori, independent with the following structure

$$\theta_j | \tau_j^2 \sim \mathcal{N}(0, \tau_j^2), \qquad \tau_j^2 \sim \mathcal{E}\left(\frac{\lambda^2}{2}\right),$$

where $\mathcal{E}$ denotes the exponential distribution with mean $2/\lambda^2$.

For shrinking the time-varying coefficients, a lasso extension was used by assigning for the standard deviations $\boldsymbol{\omega} = (\omega_1, .., \omega_k)'$ an hierarchical mixture of normal priors with an exponential mixing density. That is, each element of $\boldsymbol{\omega}$ is assumed to be, a priori, conditionally independent with

$$\omega_j | \xi_j^2 \sim \mathcal{N}(0, \xi_j^2), \qquad \xi_j^2 | \kappa \sim \mathcal{E}\left(\frac{\kappa^2}{2}\right). \tag{3.18}$$

Finally, the specification is completed by assuming a stochastic volatility specification for σ_t^2 and hyperpriors $\lambda \sim \mathcal{G}(a_1, a_2)$ and $\kappa \sim \mathcal{G}(b_1, b_2)$.

The methodology explained above decides, in an automatic fashion, whether any predictor is important for forecasting inflation and, if so, whether it has a coefficient which is constant over time or time-varying. In terms of Equation (3.17), the model has four outcomes/combinations:

- If ω_j is shrunk to 0, but θ_j is not shrunk to 0, then the model has a constant parameter on predictor j.
- If ω_j is shrunk to 0 and θ_j is shrunk to 0, then predictor j is irrelevant for forecasting inflation.

- If ω_j is not shrunk to 0, but θ_j is shrunk to 0, then the model has a small time-varying coefficient on predictor j (small since $\theta_{j,0} = 0$, that is, the coefficient starts at value 0).
- If ω_j is not shrunk to 0 and θ_j is not shrunk to 0, then we have an unrestricted time-varying coefficient on predictor j .

Posterior computation is based on a MCMC algorithm, where the block $\tilde{\boldsymbol{\theta}}_{1:T} = (\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_T)'$, with $\boldsymbol{\theta}_t = (\theta_{1,t}, ..., \theta_{k,t})$ for $t = 1, ..., T$, is drawn using the FFBS algorithm presented in Algorithm 4. Note that conditional on $\boldsymbol{\omega}$, the model becomes a Gaussian DLM, allowing for drawing the states using FFBS. For the stochastic volatility treatment, the algorithm of Kim et al. (1998) was used. For details about the full conditionals, please refer to Belmonte et al. (2014).

Is is worth pointing out that, it was assigned a Normal prior for the standard deviations ω_j, $j = 1..., k$ in order to induce sparsity, differently from the traditional Bayesian approaches, which generally assign Inverted Gamma priors for the variances ω_j^2. The choice follows the work of Frühwirth-Schnatter and Wagner (2010), where it is argued and presented evidence that the hyperparameters of the Inverted-Gamma prior strongly influence the posterior density of ω_j if the true value of ω_j^2 is close to 0 (that is, the case when the jth time-varying coefficient is irrelevant), shrinking the posterior of ω_j away from 0.

In a recent paper, Bitto and Frühwirth-Schnatter (2016) adopted a very similar approach from Belmonte et al. (2014), encouraging shrinkage in TVP models using a scaled mixture of normal prior for the standard deviation ω_j with a Gamma mixing distribution instead, which corresponds to the Normal-Gamma prior from Griffin et al. (2010) as discussed in Section 2.2.

In their work they also used the non-centered parametrization of the dynamic linear regression for shrinking. Note that the centered parametrization of the dynamic linear regression model was presented in the Equation(3.9). Using the same notation from this equation, the non-centered parametrization considered by Bitto and Frühwirth-Schnatter (2016) is as follows

$$\begin{aligned} y_t &= \boldsymbol{X}_t\boldsymbol{\beta} + \boldsymbol{X}_t\boldsymbol{W}_t^{-1/2}\tilde{\boldsymbol{\beta}}_t + \nu_t \\ \tilde{\boldsymbol{\beta}}_t &= \tilde{\boldsymbol{\beta}}_t + \tilde{\boldsymbol{\omega}}_t, \\ \tilde{\boldsymbol{\beta}}_0 &\sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{P}_0), \end{aligned} \tag{3.19}$$

for $t = 1, ..., n$, where $\boldsymbol{X}_t$ is a $(1 \times q)$ vector of regressors, $\boldsymbol{\beta} = (\beta_1, ..., \beta_q)'$ is a $(q \times 1)$ vector of fixed coefficients over time, $\tilde{\boldsymbol{\beta}}_t = (\tilde{\beta}_{1t}, ..., \tilde{\beta}_{qt})'$ is a $(q\times 1)$ vector of time-varying scaled coefficients over time, where $\tilde{\beta}_{jt} = \beta_{jt}/\omega_j$ for $j = 1, .., q$, and $\boldsymbol{W}_t = \boldsymbol{W} = \text{diag}(\omega_1^2, ..., \omega_q^2)$ is the covariance matrix from the centered parametrization, which is assumed to be diagonal and fixed over time. Hence, $\tilde{\boldsymbol{\omega}}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$.

Concerning the error variances in the measurement equation, it was considered the homoscedastic case $\boldsymbol{\nu}_t \sim \mathcal{N}(0, \sigma^2)$ for all t as well as a more flexible model specification, where $\boldsymbol{\nu}_t \sim \mathcal{N}(0, \sigma_t^2)$. To capture heteroscedasticity, they used a stochastic volatility specification as in Jacquier et al. (1994) for σ_t^2.

To avoid any scaling issues, it is assumed that all covariates, except a possible intercept, have been standardized such that for each j the average of X_{tj} over t is zero and the sample variance is equal to 1.

Note that the initial value in the non-centered parameterization is assumed to be random, i.e., $\tilde{\boldsymbol{\beta}}_0 \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{P}_0)$ rather than zero as in earlier work of Belmonte et al. (2014) (compare with the earlier Equation (3.17)). That is, it is assumed that $(\beta_{j,0}|\beta_j, \omega_j) \sim \mathcal{N}(\beta_j, \omega_j^2 P_{0,jj})$, where β_j is the fixed coefficient related to covariate $\boldsymbol{X}_j$ and ω_j^2 is the process variance governing the dynamics of the time-varying coefficient β_{jt}. Accordingly to the authors, this additional randomness avoids overshrinking of the time-varying parameters $\boldsymbol{\beta}_t$ toward $\boldsymbol{\beta}$ for the first few time points.

The rationale behind the specification (3.19) is to pull time-varying regression coefficients $\beta_{j,1}, ..., \beta_{j,n}$ toward the fixed regression coefficient β_j, if the TVP model is overfitting and the effect of the jth covariate $\boldsymbol{X}_j$ is in fact not changing over time. This requires the definition of priors on the process variances ω_j^2 that are able to shrink ω_j^2 toward the boundary value 0, but, at the same time, are flexible enough to avoid overshrinking for regression coefficients that are, actually, changing over time t and are characterized by a non-zero process variance ($\omega_j^2 \neq 0$).

As mentioned before, a very popular prior choice for the process variance ω_j^2 is the Inverse-Gamma prior distribution, which is the conjugate prior for ω_j^2 in the centered parameterization. However, following Frühwirth-Schnatter and Wagner (2010), they prefered to use Gaussian priors for the standard deviations ω_j. Differently from Belmonte et al. (2014), who have used Bayesian lasso for shrinking ω_j, it is assumed a Normal-Gamma prior for ω_j as follows

$$\omega_j|\xi_j^2 \sim \mathcal{N}(0, \xi_j^2), \qquad (\xi_j^2|a^\xi, \kappa^2) \sim \mathcal{G}(a^\xi, a^\xi\kappa^2/2), \tag{3.20}$$

where $\mathcal{G}$ denotes the Gamma distribution with mean $2/\kappa^2$ (shape-rate parametrization). For $a^\xi = 1$ the Gamma distribution reduces to the Exponential distribution and the Bayesian lasso adopted by Belmonte et al. (2014) in Equation (3.18) is a special case of (3.20).

In terms of the process variances ω_j^2, (3.20) implies that ω_j^2 follows a "double Gamma" prior

$$\omega_j|^2\xi_j^2 \sim \mathcal{G}(1/2, 1/(2\xi_j^2)), \qquad (\xi_j^2|a^\xi, \kappa^2) \sim \mathcal{G}(a^\xi, a^\xi\kappa^2/2). \tag{3.21}$$

To infer an appropriate value of κ^2 from the data, yet another layer of hierarchy is added, by assuming that the hyperparameter κ^2 is random, following again a gamma distribution

$$\kappa^2 \sim \mathcal{G}(d_1, d_2),$$

while a^ξ is supposed fixed at the value $a^\xi = 0.1$ based on simulations made by the authors. As argued by them, the prior with $a^\xi = 0.1$ showed a much more flexible shrinkage behavior than the Bayesian lasso $a^\xi = 1$. In fact, this value is the same used by Kastner (2016) as mentioned before in Section 2.2.

If ω_j^2 is shrunken toward 0, then this pulls $\beta_{j,0}$ and all subsequent values $\beta j, t$ toward the fixed regression parameter β_j. Then, it is also relevant to allow shrinkage of β_j toward 0. The also used the the Normal-Gamma prior for shrinking the fixed coefficients β_j for $j = 1, ..., q$, i.e.,

$$\beta_j|\tau_j^2 \sim \mathcal{N}(0, \tau_j^2), \qquad (\tau_j^2|a^\tau, \lambda^2) \sim \mathcal{G}(a^\tau, a^\tau\lambda^2/2). \tag{3.22}$$

Also another layer of hierarchy is added by assuming

$$\lambda^2 \sim \mathcal{G}(e_1, e_2),$$

while assuming $a^\tau = 0.01$. Closed forms for the densities of $\beta_{j,t}$ and β_j can be inferred using the Normal-Gamma marginal density presented in the Equation(2.21) from Section 2.2.

To carry out Bayesian inference under the shrinkage priors (3.20) and (3.22), it was developed efficient schemes for full conditional Markov chain Monte Carlo (MCMC) sampling based on the ancillarity sufficiency interweaving strategy (ASIS) introduced by Yu and Meng (2011). For details on what they named efficient full conditional Gibbs sampling for sparse TVP models, please refer to Section 3.1 from Bitto and Frühwirth-Schnatter (2016).

Finally, the proposed model and the sampling strategy were applied to two real world data set: the first application was an inflation modeling using the same data analyzed in Belmonte et al. (2014) and the second application was a TVP Cholesky stochastic volatility modeling of 29 returns from German Stock Index DAX.

3.3.3 Time-varying dimension models

In the previous sections we have presented many approaches for inducing sparsity or parsimony in time-varying parameter models. However, neither of these methods treated the sparsity structure as time-varying. From now on, we turn our attention to approaches where the subset of relevant variables varies over time.

A distinguish reference about shrinkage in TVP models is Chan et al. (2012), whose focus was also inflation forecasting. According to the authors, TVP models are parameter-rich and risk-overfitting unless the dimension of the model is small. Therefore, through a Bayesian approach considering the estimation of state space models subject to equality restrictions on the states, they have proposed a method where the dimension of the model can change over time, allowing for an automatically choice over different parsimonious representations.

Their approach belongs to the category of dynamic mixture models, which will be discussed in Section 3.3.4 and was named *time-varying dimension* (TVD) models. The model specification is given below.

Consider the dynamic linear regression from Equation (3.9) with a slight modification, i.e.,

$$\begin{aligned} y_t &= \boldsymbol{X}_t \boldsymbol{\beta}_t + \nu_t, \\ \boldsymbol{\beta}_{t+1} &= \boldsymbol{\beta}_t + \boldsymbol{\omega}_t, \end{aligned} \tag{3.23}$$

for each $t = 1, ..., n$, where $\boldsymbol{X}_t$ is the $(1 \times q)$ vector of regressors, $\boldsymbol{\beta}_t' = (\beta_{1t}, ..., \beta_{qt})$ is the $(q \times 1)$ vector of coefficients, $\nu_t \sim \mathcal{N}(0, V_t)$ and $\boldsymbol{\omega}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{W}_t)$, with errors ν_t and $\boldsymbol{\omega}_t$ independent of each other.

The dynamic mixture model of Gerlach et al. (2000) adds to (3.24) the assumption that the values of the system matrices $\boldsymbol{X}_t$ and $\boldsymbol{W}_t$ and the variance V_t are determined, up to a set of unknown parameters, by the value of the vector $\boldsymbol{K}_t$, where $\boldsymbol{K}_{1:n} = (\boldsymbol{K}_1, ..., \boldsymbol{K}_n)$ is a sequence of random vectors that are Markov, i.e., $p(\boldsymbol{K}_t | \boldsymbol{K}_{1:t-1}) = p(\boldsymbol{K}_t | \boldsymbol{K}_{t-1})$, for $t = 2, ..., n$.

The great contribution of Gerlach et al. (2000) was developing an efficient algorithm for posterior simulation for this class of models. The efficiency gains occur since the states $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_n)$ are integrated out and $\boldsymbol{K}_{1:n}$ is drawn unconditionally. A simple alternative algorithm would involve drawing from the posterior of $\boldsymbol{K}_{1:n}$, conditional on the states $\boldsymbol{\beta}_{1:n}$, and then the posterior of $\boldsymbol{\beta}_{1:n}$, conditional on $\boldsymbol{K}_{1:n}$. Such a strategy was shown to produce a chain of draws which is very slow to mix. In Chan et al. (2012), the algorithm of Gerlach et al. (2000) was used to draw the posterior of $\boldsymbol{K}_{1:n}$, considering three different ways $\boldsymbol{K}_t$ can enter the system matrices so as to yield a TVD model.

The first TVD model adapts the approach of Gerlach et al. (2000) in a particular way such that $\boldsymbol{\beta}_t$ remains a q-dimensional vector at all times, but there is a sense in which the dimension of the model can change over time. This can be done by allowing for explanatory variables to be included/excluded from the likelihood function depending on $\boldsymbol{K}_t$.

The basic idea can be illustrated as follows. Suppose $q = 1$ and $X_t = K_t z_t$, where z_t is an explanatory variable and $K_t \in \{0, 1\}$. If $K_t = 0$, then z_t does not enter the likelihood, and the coefficient β_t does not enter the model. However, if $K_s = 1$, then the coefficient β_s enters the model. Thus the dimension of the model is different from time t than at time s. It is worth stressing that, if $K_t = 0$, then β_t does not enter the likelihood, but the dimension of the state β remains the same.

To make this idea clear, suppose we have a model with the response y depending on a vector of parameters $\boldsymbol{\theta}$ which are partitioned as $\boldsymbol{\theta} = (\boldsymbol{\phi}, \boldsymbol{\gamma})$. Then the prior is $p(\boldsymbol{\theta}) = p(\boldsymbol{\gamma})p(\boldsymbol{\phi}|\boldsymbol{\gamma})$ and the likelihood is $\ell(y|\boldsymbol{\theta})$.

Consider a second model which imposes the restriction $\boldsymbol{\phi} = \mathbf{0}$. Instead of directly imposing this restriction, consider what happens if we impose the restriction that $\boldsymbol{\phi}$ does not enter the likelihood. That is, the likelihood for the second model would be $\ell(y|\boldsymbol{\theta}) = \ell(y|\boldsymbol{\gamma})$ and its posterior would be

$$p(\boldsymbol{\theta}|y) = \frac{\ell(y|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int \ell(y|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}} = \frac{\ell(y|\boldsymbol{\gamma})p(\boldsymbol{\gamma})}{\int \ell(y|\boldsymbol{\gamma})p(\boldsymbol{\theta})d\boldsymbol{\theta}} p(\boldsymbol{\phi}|\boldsymbol{\gamma}) = p(\boldsymbol{\gamma}|y)p(\boldsymbol{\phi}|\boldsymbol{\gamma}).$$

Since $p(\boldsymbol{\phi}|\boldsymbol{\gamma})$ integrates to one (or assigns a point mass to $\boldsymbol{\phi} = \mathbf{0}$) integrating $p(\boldsymbol{\theta}|y)$ with respect to $\boldsymbol{\phi}$ provides a valid posterior for the second model and the integral $\int \ell(y|\boldsymbol{\gamma})p(\boldsymbol{\theta})d\boldsymbol{\theta}$ would result in the correct marginal likelihood. This is the strategy which underlies and justifies the first approach.

In summary, assuming the same structure for the errors and the design matrix as in (3.23), the first TVD model is then

$$\begin{aligned} y_t &= \boldsymbol{X}_t \boldsymbol{M}_t \boldsymbol{\beta}_t + \nu_t, \\ \boldsymbol{\beta}_{t+1} &= \boldsymbol{\beta}_t + \boldsymbol{\omega}_t, \end{aligned} \tag{3.24}$$

for $t = 1, .., n$, where $\boldsymbol{M}_t = \mathrm{diag}(K_{1,t}, ..., K_{q,t})$ and $K_{j,t} \in \{0, 1\}$, $j = 1, ..., q$.

The second TVD model is based on what is known in Bayesian Vector Autoregression (BVAR) literature as the Minnesota prior, which was first proposed by Litterman et al. (1979). The model is a follows (assuming the same structure for the errors and the design matrix as in (3.23))

$$\begin{aligned} y_t &= \boldsymbol{X}_t \boldsymbol{\beta}_t + \nu_t, \\ \boldsymbol{\beta}_{t+1} &= \boldsymbol{M}\boldsymbol{\beta}_t + (\boldsymbol{I} - \boldsymbol{M})\bar{\boldsymbol{\beta}} + \boldsymbol{\omega}_t, \end{aligned}$$

for $t = 1, .., n$, where $\boldsymbol{M}$ is a $(q \times q)$ matrix and $\bar{\boldsymbol{\beta}}$ is a $q \times 1$ vector. In particular, the authors set $\boldsymbol{M} = \boldsymbol{M}_t = \mathrm{diag}(K_{1,t}, ..., K_{q,t})$, $\bar{\boldsymbol{\beta}} = \mathbf{0}$ and $W_t = M_t W$. That is, if $K_{j,t} = 1$, the jth coefficient evolves according to a random walk as in standard TVP regression. But if $K_{j,t} = 0$, then coefficient is set to zero, and the dimension of the model reduces.

The third TVD model formally reduces the dimension of the state vector $\boldsymbol{\beta}_t$ as it specified as follows (again, assuming the same structure for the errors and the design matrix as in (3.23))

$$\begin{aligned} y_t &= \boldsymbol{X}_t\boldsymbol{\beta}_t + \nu_t, \\ \boldsymbol{\beta}_{t+1} &= \boldsymbol{M}\boldsymbol{\theta}_{t+1} + \boldsymbol{\omega}_t, \\ \boldsymbol{\theta}_{t+1} &= \boldsymbol{\theta}_t + \boldsymbol{u}_t, \end{aligned} \tag{3.25}$$

for $t = 1, .., n$, where $\boldsymbol{u}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{R})$ and $\boldsymbol{u}_t$ is independent from the other errors in the model. The third TVD model can be constructed by specifying $\boldsymbol{M}$ and $\boldsymbol{W}_t$ to be exactly as in the second TVD model. That is, $\boldsymbol{\omega}_t \sim \mathcal{N}(0, M_t W)$ and $\boldsymbol{M} = \boldsymbol{M}_t = \text{diag}(K_{1,t}, ..., K_{q,t})$.

Note that $\boldsymbol{\theta}_t$ can potentially be of a lower dimension that $\boldsymbol{\beta}_t$ and that the two last equations from (3.25) can be rewritten as

$$\boldsymbol{\beta}_{t+1} = \boldsymbol{\beta}_t + \boldsymbol{v}_t,$$

where $\boldsymbol{v}_t = \boldsymbol{M}\boldsymbol{u}_t + \boldsymbol{\omega}_t - \boldsymbol{\omega}_{t-1}$, that is, a moving average (MA) structure.

Note that the first and the third model has similar properties. In order to understand these properties consider the case where a coefficient drops out of the model for h periods and then reenters it, i.e.,

$$K_{j,t-1} = 1, \quad K_{j,t} = K_{j,t+1} = ... = K_{j,t+h-1} = 0, \quad K_{j,t+h} = 1.$$

In this case, the first TVD model, supposing that $\boldsymbol{W}_t = \boldsymbol{W}$, implies $\mathbb{E}(\boldsymbol{\beta}_{t+h}) = \boldsymbol{\beta}_{t-1}$ and $\text{Var}(\boldsymbol{\beta}_{t+h}) = h\boldsymbol{W}$ while the second model implies $\mathbb{E}(\boldsymbol{\beta}_{t+h}) = \boldsymbol{0}$ and $\text{Var}(\boldsymbol{\beta}_{t+h}) = \boldsymbol{W}$. The third model has properties closer to those of the first approach and yields $\mathbb{E}(\boldsymbol{\beta}_{t+h}) = \boldsymbol{\beta}_{t-1}$ and $\text{Var}(\boldsymbol{\beta}_{t+h}) = h\boldsymbol{R} + \boldsymbol{W}$ if $\boldsymbol{M}$ is a square matrix.

The three different TVD models can be implemented with any choice of $\boldsymbol{K}_t$. However, the approach can become computationally demanding if the dimension of $\boldsymbol{K}_t$ is large. If one let $\boldsymbol{K}_t$ be a vector of q dummy variables controlling whether each regressor out of q regressors is included or excluded in the model at time t, then there would be 2^q values $\boldsymbol{K}_t$ could take and the computational demands will be high unless q is small.

Therefore, to handle the computational problem they have limited the values that $\boldsymbol{K}_t$ could take, precisely, $\boldsymbol{K}_t$ could only take values in

$$\mathcal{I} = \{(0, 0, ..., 0), (1, 0, ..., 0), (0, 1, 0, ..., 0), (0, 0, ..., 0, 1), (1, 1, ..., 1)\},$$

which totals $q + 2$ possible values.

In addition, they imposed a Markov hierarchical prior for $\boldsymbol{K}_{1:n}$ as

$$\begin{aligned} &p(K_{t+1} = i | K_t = i) = c, && i \in \mathcal{I} \\ &p(K_{t+1} = j | K_t = i) = \frac{1-c}{q+1}, && i \neq j,\ i, j \in \mathcal{I}, \end{aligned} \tag{3.26}$$

for $t = 1, ..., n-1$. This prior expresses the belief that, with probability c the model will stay with its current set of explanatory variables and with probability $1 - c$ it will switch to a new model. It was assumed that c follows a Beta distribution.

For the analyzed data (core inflation as measured by the Personal Consumption Expenditure (PCE) deflator for 1962Q1 through 2008Q3), the authors set the following hyperparameters for the Beta distribution: $c \sim \mathcal{B}(4.54, 10)$, such as $\mathbb{E}(c) = 0.31$.

One must specify the initial values $K_{1,1}, ..., K_{q,1}$, because the Markov hierarchical prior in (3.26) is for $(\boldsymbol{K}_2, ..., \boldsymbol{K}_n)$. Thus, for the initial values, it was assumed that each $K_{j,1}$ is independent of each other and follows, a priori, a Bernouilli distribution with $p(K_{j,1} = 1) = b_j$, for $j = 1, ..., q$, where $b_j \sim \mathcal{B}(1.5, 1.5)$ such that $\mathbb{E}(b_j) = 0.5$. For further details about each TVD model and the real data application, please refer to the article.

In all the TVD models, MCMC methods was used to draw the posteriors, including the drawing of $\boldsymbol{K}_{1:n}$, where the algorithm ofGerlach et al. (2000) was used. For details on the posterior computations please refer to Section 2.3 from Chan et al. (2012) as well as the online appendix available in https://sites.google.com/site/garykoop/research.

3.3.4 Efficient Bayesian inference for dynamic mixture models

In this section, we present in details the dynamic mixture model approach proposed by Gerlach et al. (2000). This approach is used in the model that will be presented later in Chapter 5.

Consider the univariate Gaussian DLM from Definition 3.2 with a slight modification, i.e., allowing for time-varying intercepts in both observation and state equation as

$$\begin{aligned} y_t &= f_t + \boldsymbol{F}_t' \boldsymbol{\theta}_t + \gamma_t u_t, \qquad u_t \sim \mathcal{N}(0, 1) \\ \boldsymbol{\theta}_t &= \boldsymbol{g}_t + \boldsymbol{G}_t \boldsymbol{\theta}_{t-1} + \boldsymbol{\Gamma} \boldsymbol{v}_t, \qquad \boldsymbol{v}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}), \end{aligned} \tag{3.27}$$

for $t = 1, .., n$, where $\boldsymbol{\theta}_t$ is a q-dimensional vector of states, u_t and $\boldsymbol{v}_t$ are independent and standard Normal distributed, and f_t, $\boldsymbol{F}_t'$, γ_t, $\boldsymbol{g}_t$, $\boldsymbol{G}_t$ and $\boldsymbol{\Gamma}_t$ may all depend on the vector Markov $\boldsymbol{K}_t$ and on a vector of parameters $\boldsymbol{\Phi}$. This makes observations y_t mixture of normals.

Note that $\boldsymbol{K}_{1:n} = (\boldsymbol{K}_1, ..., \boldsymbol{K}_n)$ is a sequence of random vectors that are Markov as for $t = 2, ..., n$

$$p(\boldsymbol{K}_t | \boldsymbol{K}_{1:t-1}) = p(\boldsymbol{K}_t | \boldsymbol{K}_{t-1}).$$

The sampling scheme proposed generates $\boldsymbol{K}_t$ from density $p(\boldsymbol{K}_t | y_{1:n}, \boldsymbol{K}_{s \neq t})$ for $t = 1, ..., n$ without conditioning on the states $\boldsymbol{\theta}_{1:n}$. The crucial thing is to notice that

$$\begin{aligned} p(\boldsymbol{K}_t | y_{1:n}, \boldsymbol{K}_{s \neq t} &\propto p(y_{1:n} | \boldsymbol{K}_{1:n}) p(\boldsymbol{K}_t | \boldsymbol{K}_{s \neq t}) \\ &\propto p(y_{t+1:n} | y_{1:t}, \boldsymbol{K}_{1:n}) p(y_t | y_{1:t-1}, \boldsymbol{K}_{1:t}) p(\boldsymbol{K}_t | \boldsymbol{K}_{s \neq t}), \end{aligned} \tag{3.28}$$

where the dependence on the parameters $\boldsymbol{\Phi}$ has been suppressed for convenience.

For each value of $\boldsymbol{K}_t$ the right size of (3.28) is evaluated as follows. The term $p(\boldsymbol{K}_t | \boldsymbol{K}_{s \neq t})$ is obtained from the prior, for example the prior used by Chan et al. (2012) which is presented in the Equation (3.26). The term $p(y_t | y_{1:t-1}, \boldsymbol{K}_{1:t})$ is obtained from $p(\boldsymbol{\theta}_{t-1} | y_{1:t-1}, \boldsymbol{K}_{1:t-1})$, i.e., from the the filtering distribution, using one step of the Kalman filter presented in Proposition 3.3.

Obtaining the term $p(y_{t+1:n} | y_{1:t}, \boldsymbol{K}_{1:n})$ is the crucial innovation of the algorithm of Gerlach et al. (2000). Traditional sampling algorithms use $n - t + 1$ steps of the Kalman filter given the current values of $\boldsymbol{K}_{t,n}$ to obtain the term $p(y_{t+1:n} | y_{1:t}, \boldsymbol{K}_{1:n})$. Therefore, it requires $\mathcal{O}(n)$ operations to generate each $\boldsymbol{K}_t$, and hence $\mathcal{O}(n^2)$ operations to generate $\boldsymbol{K}_{1:n}$. Nevertheless, in the proposed algorithm the term $p(y_{t+1:n} | y_{1:t}, \boldsymbol{K}_{1:n})$ is obtained in one step after an initial set of backward recursions. This reduces the number of operations required to generate the complete vector $\boldsymbol{K}_{1:n}$ to $\mathcal{O}(n)$.

Before giving the efficient method for generating $\boldsymbol{K}_{1:n}$, we are going to state several preliminary lemmas, whose proofs can be found in the Appendix of Gerlach et al. (2000). All of the lemmas refer to the univariate Gaussian DLM (3.27).

Lemma 3.1. *Let $r_{t+1} = \mathrm{Var}(y_{t+1}|\boldsymbol{\theta}_t, \boldsymbol{K}_{1:t+1})$. Then, the following hold:*

$$\mathbb{E}(y_{t+1}|\boldsymbol{\theta}_t, \boldsymbol{K}_{1:t+1}) = f_{t+1} + \boldsymbol{F}'_{t+1}(\boldsymbol{g}_{t+1} + \boldsymbol{G}_{t+1}\boldsymbol{\theta}_t),$$

$$r_{t+1} = \boldsymbol{F}'_{t+1}\boldsymbol{\Gamma}_{t+1}\boldsymbol{\Gamma}'_{t+1}\boldsymbol{F}_{t+1} + \gamma^2_{t+1},$$

and

$$\mathbb{E}(\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t, y_{t+1}, \boldsymbol{K}_{1:n}) = \boldsymbol{a}_{t+1} + \boldsymbol{A}_{t+1}\boldsymbol{\theta}_t + \boldsymbol{B}_{t+1}y_{t+1},$$

$$\mathrm{Var}(\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t, y_{t+1}, \boldsymbol{K}_{1:n}) = \boldsymbol{C}_{t+1}\boldsymbol{C}'_{t+1},$$

where

$$\boldsymbol{a}_{t+1} = (\boldsymbol{I} - \boldsymbol{B}_{t+1}\boldsymbol{F}'_{t+1})\boldsymbol{g}_t - \boldsymbol{B}_{t+1}f_t,$$

$$\boldsymbol{A}_{t+1} = (\boldsymbol{I} - \boldsymbol{B}_{t+1}\boldsymbol{F}'_{t+1})\boldsymbol{G}_{t+1},$$

$$\boldsymbol{B}_{t+1} = \boldsymbol{\Gamma}_{t+1}\boldsymbol{\Gamma}'_{t+1}\boldsymbol{F}_{t+1}r^{-1}_{t+1},$$

$$\boldsymbol{C}_{t+1}\boldsymbol{C}'_{t+1} = \boldsymbol{\Gamma}_{t+1}(\boldsymbol{I} - \boldsymbol{\Gamma}'_{t+1}\boldsymbol{F}_{t+1}r^{-1}_{t+1}\boldsymbol{F}'_{t+1}\boldsymbol{\Gamma}_{t+1})\boldsymbol{\Gamma}_{t+1},$$

It is straightforward to factor the expression on the right side of the last equality to get a matrix $\boldsymbol{C}_{t+1}$ that either is null or has full column rank. Then, we can write

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{a}_{t+1} + \boldsymbol{A}_{t+1} + \boldsymbol{B}_{t+1}y_{t+1} + \boldsymbol{C}_{t+1}\boldsymbol{\xi}_{t+1},$$

where $\boldsymbol{\xi}_{t+1} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ and is independent of $\boldsymbol{\theta}_t$ and y_{t+1}, conditional on $\boldsymbol{K}_{1:n}$.

Lemma 3.2. *For $t = 1, ..., n-1$, the density $p(y_{t+1:n}|\boldsymbol{\theta}_t, \boldsymbol{K}_{1:n})$ is independent of $\boldsymbol{K}_{1:t}$ and can be expressed as*

$$p(y_{t+1:n}|\boldsymbol{\theta}_t, \boldsymbol{K}_{1:n}) \propto \exp\left\{-\frac{1}{2}(\boldsymbol{\theta}_t\boldsymbol{\Omega}_t(\boldsymbol{\theta}_t - 2\boldsymbol{\mu}'_t\boldsymbol{\theta}_t)\right\},$$

where the terms $\boldsymbol{\Omega}_t$ and $\boldsymbol{\mu}_t$ are computed recursively starting from

$$\boldsymbol{\Omega}_n = \boldsymbol{0}, \qquad \boldsymbol{\mu}_n = \boldsymbol{0}$$

and moving backward

$$\boldsymbol{\Omega}_t = \boldsymbol{A}'_{t+1}(\boldsymbol{\Omega}_{t+1} - \boldsymbol{\Omega}_{t+1}\boldsymbol{C}_{t+1}\boldsymbol{D}^{-1}_{t+1}\boldsymbol{C}'_{t+1}\boldsymbol{\Omega}_{t+1})\boldsymbol{A}_{t+1} + \boldsymbol{G}'_{t+1}\boldsymbol{F}_{t+1}r^{-1}_{t+1}\boldsymbol{F}'_{t+1}\boldsymbol{G}_{t+1},$$

$$\boldsymbol{\mu}_t = \boldsymbol{A}'_{t+1}(\boldsymbol{I} - \boldsymbol{\Omega}_{t+1}\boldsymbol{C}_{t+1}\boldsymbol{D}^{-1}_{t+1}\boldsymbol{C}'_{t+1})(\boldsymbol{\mu}_{t+1} - \boldsymbol{\Omega}_{t+1}(\boldsymbol{a}_{t+1} + \boldsymbol{B}_{t+1}y_{t+1})) + \boldsymbol{G}'_{t+1}\boldsymbol{F}_{t+1}r^{-1}_{t+1}(y_{t+1} - f_{t+1} - \boldsymbol{F}'_{t+1}\boldsymbol{g}_{t+1}),$$

where

$$\boldsymbol{D}_{t+1} = \boldsymbol{C}'_{t+1}\boldsymbol{\Omega}_{t+1}\boldsymbol{C}_{t+1} + \boldsymbol{I}.$$

Lemma 3.3. *Let* $\boldsymbol{m}_t = \mathbb{E}(\boldsymbol{\theta}_t|y_{1:t}, \boldsymbol{K}_{1:n})$*,* $\boldsymbol{V}_t = \mathrm{Var}(\boldsymbol{\theta}_t|y_{1:t}, \boldsymbol{K}_{1:n})$ *and* $R_t = \mathrm{Var}(y_t|y_{1:t-1}, \boldsymbol{K}_{1:n})$*. The Kalman filter for the model* (3.27) *is given by*

$$\boldsymbol{R}_t = \boldsymbol{F}'_t\boldsymbol{G}_t\boldsymbol{V}_{t-1}\boldsymbol{G}'_t\boldsymbol{F}_t + \boldsymbol{F}'_t\boldsymbol{\Gamma}_t\boldsymbol{\Gamma}'_t\boldsymbol{F}_t + \gamma_t^2,$$

$$\boldsymbol{m}_t = (\boldsymbol{I} - \boldsymbol{J}_t\boldsymbol{F}'_t)(\boldsymbol{g}_t + \boldsymbol{G}_t\boldsymbol{m}_{t-1}) + \boldsymbol{J}_t(y_t - f_t),$$

$$\boldsymbol{V}_t = \boldsymbol{G}_t\boldsymbol{V}_{t-1}\boldsymbol{G}'_t + \boldsymbol{\Gamma}_t\boldsymbol{\Gamma}'_t - \boldsymbol{J}_t\boldsymbol{J}'_tR_t,$$

where

$$\boldsymbol{J}_t = [\boldsymbol{G}_t\boldsymbol{V}_{t-1}\boldsymbol{G}'_t\boldsymbol{F}_t + \boldsymbol{\Gamma}_t\boldsymbol{\Gamma}'_t\boldsymbol{F}_t]/R_t.$$

The conditional density $p(y_t|y_{1:t-1}, \boldsymbol{K}_{1:t}) \propto R_t^{-1}\exp\left\{-\frac{1}{2R_t}(y_t - f_t - \boldsymbol{F}'_t(\boldsymbol{g}_t + \boldsymbol{G}_t\boldsymbol{m}_{t-1}))^2\right\}$*.*

We can write $\boldsymbol{V}_t = \boldsymbol{T}_t\boldsymbol{T}'_t$*, where the matrix* $\boldsymbol{T}_t$ *either has full column rank if* $\boldsymbol{V}_t \neq \boldsymbol{0}$ *or is null if* $\boldsymbol{V}_t = \boldsymbol{0}$*. Conditional on* $\boldsymbol{K}_{1:n}$*, we can express* $\boldsymbol{\theta}_t$ *as*

$$\boldsymbol{\theta}_t = \boldsymbol{m}_t + \boldsymbol{T}_t\boldsymbol{\xi}_t,$$

where $\boldsymbol{\xi}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ *and is independent of* $y_{1:t}$*.*

The next Lemma uses Lemma 3.3 to efficiently evaluate the factor $p(y_{t+1:n}|y_{1:t}, \boldsymbol{K}_{1:n})$.

Lemma 3.4. *Using the results of Lemma 3.3, it follows that*

$$\begin{aligned} p(y_{t+1:n}|y_{1:t}, \boldsymbol{K}_{1:n}) &= \int p(y_{t+1:n}|\boldsymbol{\theta}_t, \boldsymbol{K}_{t+1:n})p(\boldsymbol{\xi}_t|\boldsymbol{K}_{1:t})d\boldsymbol{\xi}_t \\ &\propto |\boldsymbol{T}'_t\boldsymbol{\Omega}_t\boldsymbol{T}_t + \boldsymbol{I}|^{-1/2}\exp\left\{-\frac{1}{2}\left(\boldsymbol{m}'_t\boldsymbol{\Omega}_t\boldsymbol{m}_t - 2\boldsymbol{\mu}'_t\boldsymbol{m}_t - \boldsymbol{\phi}'_t(\boldsymbol{T}'_t\boldsymbol{\Omega}_t\boldsymbol{T}_t + \boldsymbol{I})^{-1}\boldsymbol{\phi}_t\right)\right\}, \end{aligned}$$

where $\boldsymbol{\phi}_t = \boldsymbol{T}'_t(\boldsymbol{\mu}_t - \boldsymbol{\Omega}_t\boldsymbol{m}_t)$*.*

The recursion for generating $\boldsymbol{K}_{1:n}$ in $\mathcal{O}(n)$ operations is now given.

3.3.5 Time-varying sparsity in dynamic regression models

In this section we present the approach of Kalli and Griffin (2014), a novel Bayesian method for inference in dynamic regression models where both the values of the regression coefficients and the importance of the variables are allowed to change over time. In order to allow for a time-varying sparsity, an extension of the Normal-Gamma prior (see, e.g., Griffin et al. (2010)) for dynamic regression was developed which allows the shrinkage of the regression coefficients to suitably change over time.

In regression models with a large number of predictors, it is common to assume that only a subset of them is important for prediction. In the context of dynamic regression it is reasonable

Algorithm 6: The algorithm of Gerlach et al. (2000) for dynamic mixture models

```
1. Given the current value of $\boldsymbol{K}_{1:n}$, calculate $\boldsymbol{\Omega}_t$ and $\boldsymbol{\mu}_t$ for $t = n-1, ..., 1$, using the
   recursions in Lemma 3.2.
2. Given $\mathbb{E}(\boldsymbol{\theta}_0)$ and $\mathrm{Var}(\boldsymbol{\theta}_0)$, perform the following for $t = 1, ...n$:
(a) Obtain $R_t$, $\boldsymbol{m}_t$ and $\boldsymbol{V}_t$ from $\boldsymbol{m}_{t-1}$ and $\boldsymbol{V}_{t-1}$ as in Lemma 3.3;
(b) Obtain $p(y_t|y_{1:t-1}, \boldsymbol{K}_{1:t})$ as in Lemma 3.3 and $p(y_{t+1:n}|y_{1:t}, \boldsymbol{K}_{1:n})$ as in Lemma 3.4;
(c) Obtain $p(\boldsymbol{K}_t|y_{1:n}\boldsymbol{K}_{s\neq t})$ for all values of $\boldsymbol{K}_t$ by normalization of

        $p(\boldsymbol{K}_t|y_{1:n}, \boldsymbol{K}_{s\neq t} \propto p(y_{1:n}|\boldsymbol{K}_{1:n})p(\boldsymbol{K}_t|\boldsymbol{K}_{s\neq t})$
        $\propto p(y_{t+1:n}|y_{1:t}, \boldsymbol{K}_{1:n})p(y_t|y_{1:t-1}, \boldsymbol{K}_{1:t})p(\boldsymbol{K}_t|\boldsymbol{K}_{s\neq t}).$

Then, draw $\boldsymbol{K}_t$.
(d) Update $\boldsymbol{m}_t$ and $\boldsymbol{V}_t$ as in Lemma 3.3, using the generated value of $\boldsymbol{K}_t$.
```

to assume that these subsets change over time. This assumption can be expressed by defining a stochastic process for the coefficients as follows. Consider the observation equation

$$y_t = \sum_{j=0}^{q} X_{j,t}\beta_{j,t} + \epsilon_t, \qquad \epsilon_t \sim \mathcal{N}(0, \sigma_t^2) \tag{3.29}$$

for $t = 1, ..., n$, where $\beta_{j,t}$ is the coefficient associated with the jth regressor in time t, $X_{j,t}$. An intercept is allowed by defining $X_{0,t} = 1$ for all t. Time-varying sparsity is allowed by giving independent Normal-Gamma autoregressive (NGAR) process priors to the time series of regression coefficients $\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_q$. That is, for $t = 2, ..., n$ and $j = 1, .., q$,

$$\beta_{j,t} = \sqrt{\frac{\psi_{j,t}}{\psi_{j,t-1}}}\varphi_j\beta_{j,t-1} + \eta_{i,t}, \qquad \eta_{j,t}|\psi_{j,t} \sim \mathcal{N}(0, (1-\varphi_j^2)\psi_{j,t}), \tag{3.30}$$

with $\beta_{j,1} \sim \mathcal{N}(0, \psi_{j,1})$. Note that this is a Normal AR(1) process conditional on $\boldsymbol{\psi}_j = (\psi_{j,1}, ..., \psi_{j,n})$, where $\psi_{j,t}$ follows a first order gamma autoregressive (GAR) process (see, e.g., Gourieroux and Jasiak (2006)), which is specified using latent variables $\kappa_{j,1}, ..., \kappa_{j,t-1}$ by the recursion

$$\begin{aligned} \psi_{j,t}|\kappa_{j,t-1} &\sim \mathcal{G}\left(\lambda_j + \kappa_{j,t-1}, \frac{\lambda_j}{\mu_j(1-\rho_j)}\right), \\ \kappa_{j,t-1}|\psi_{j,t-1} &\sim \mathcal{P}\left(\frac{\rho_j\lambda_j\psi_{j,t-1}}{\mu_j(1-\rho_j)}\right), \end{aligned} \tag{3.31}$$

for $t = 2, ..., n$, with $\psi_{j,1} \sim \mathcal{G}(\lambda_j, \lambda_j/\mu_j)$. In this definition, $\mathcal{G}(a, b)$ denotes the Gamma distribution with mean a/b and $\mathcal{P}$ denotes the Poisson distribution.

The NGAR process specified by equations (3.30) and (3.31) and notated as NGAR$(\lambda_j, \mu_j, \varphi_j, \rho_j)$ can also be represented as the product of the following two independent stationarity stochastic processes: $\boldsymbol{\psi}_j = (\psi_{j,1}, ..., \psi_{j,n})$ and $\boldsymbol{\phi}_j = (\phi_{j,1}, ..., \phi_{j,n})$. That is, for $t = 1, ..., n$,

$$\beta_{j,t} = \sqrt{\psi_{j,t}}\phi_{j,t},$$

where $\boldsymbol{\phi}$ is generated by from an AR(1) process with autocorrelation parameter φ_j such that $\boldsymbol{\phi}$ has the standard Normal as its stationarity distribution, i.e.,

$$\phi_{j,t} = \varphi_j\phi_{j-1} + \xi_j, \qquad \xi_j \sim \mathcal{N}(0, (1-\varphi^2)),$$

and $\boldsymbol{\psi}_j$ is generated from a GAR process with stationarity distribution $\mathcal{G}(\lambda_j, \lambda_j/\mu_j)$ as in (3.31).

The process $\boldsymbol{\beta}_j = (\beta_{j,1}, ..., \beta_{j,n})$ is stationary and has a Normal-Gamma stationarity distribution since $\boldsymbol{\psi}_j$ and $\boldsymbol{\phi}_j$ are independent and stationarity (assuming that $|\varphi_j| < 1$). The unconditional variance is $\text{Var}(\boldsymbol{\beta}_j) = \mu_j$ and the excess kurtosis $3/\lambda_j$.

The NGAR process prior is similar to the Normal-Gamma prior discussed in Section 2.2.2 with $\beta_{j,t}|\psi_{j,t}$ following a Normal distribution with mean zero and variance $\psi_{j,t}$ following a Gamma distribution $\psi_{j,t} \sim \mathcal{G}(\lambda_j, \lambda_j/\mu_j)$, marginally. Therefore, $\psi_{j,t}$ plays the role of relevance of jth regressor at time t. Small values of the conditional variance $\psi_{j,t}$ leads to greater shrinkage of the coefficient $\beta_{j,t}$.

For a fixed prior mean μ_j, as the value of the shrinkage parameter λ_j decreases, more prior mass for $\beta_{j,t}$ is placed close to zero and so the process $\boldsymbol{\beta}_j$ tends to spend a greater proportion of time close to zero. That is, the parameter λ_j controls the proportion of time that the regression coefficient spends close to zero, where smaller values of λ_j lead to "spikier" processes of $\psi_{j,t}$ and $\beta_{j,t}$, i.e., favor rapid changes from small to large values.

The autocorrelation parameter ρ_j controls the dependence between $\psi_{j,t}$ and $\psi_{j,t-1}$, while the autocorrelation parameter φ_j controls the dependence between $\beta_{j,t}$ and $\beta_{j,t-1}$ conditional on the $\boldsymbol{\psi}_j = (\psi_{j,1}, ..., \psi_{j,n})$ process.

Hyperpriors for the parameters of the NGAR process prior are described as follows. Parameter μ_j acts as an overall relevance parameter for the jth coefficient since it controls the unconditional variance of $\beta_{j,t}$. If μ_j is small, then $\beta_{j,t}$ will be close to zero for all t. Therefore, the authors proposed the following hierarchical prior for $\mu_1, ..., \mu_q$,

$$\mu_j \sim \mathcal{G}(\lambda^*, \lambda^*/mu^*), \qquad \lambda^* \sim \mathcal{E}(1/s^*), \qquad \mu^* \propto (\mu^* + 2b^*)^{-3},$$

where $\mathcal{E}(x)$ denotes the Exponential distribution with mean $1/x$ so that the hyperparameter s^* is the prior mean of λ^* and λ^* is the shrinkage parameter of μ_j. Smaller values of λ^* indicates that more μ_j's are close to zero. This introduces a second level of sparsity at the level of the regressors rather than the time-varying coefficients. The hyperparameter μ^* has a heavy tailed prior distribution with prior mean b^*, which is given a value suitable for the spread of the regression coefficients in the particular application.

The sparsity parameter λ_j is given the prior

$$\lambda_j \propto \lambda_j (0.5 + \lambda_j)^{-4},$$

which is a heavy tailed prior with values around 1. This centers the prior over the lasso cases (λ_j=1).

According to the authors, the flexibility of the NGAR process can lead to overfitting when the values of the autocorrelation parameters ρ_j and φ_j are small. Thus, it was assigned Beta distribution priors with a high prior mean

$$\rho_j \sim \mathcal{B}(77.6, 2.4), \qquad \varphi_j \sim \mathcal{B}(77.6, 2.4),$$

which gives a prior mean of 0.97 with most mass over 0.9 and implies that the processes for the regression coefficients $\boldsymbol{\beta}_j$ and for the relevances $\boldsymbol{\psi}_j$ are strongly autocorrelated. These priors excludes models which allow the regression coefficients to rapidly change over time and lead to overfitting.

Finally, it was assumed a stochastic volatility specification for the variances $\sigma_1^2, ..., \sigma_n^2$ from equation (3.29), where $\sigma_1^2, ..., \sigma_n^2$ follow a first order GAR process. The hyperparameter settings for the volatilities as well as the MCMC methods to fit the dynamic regression with a NGAR process as described in (3.30) and (3.31) can be found in Kalli and Griffin (2014).

The NGAR regression model was applied to both equity premium and inflation datasets and the predictive performance was compared with other sparsity inducing methods, such as those proposed by Belmonte et al. (2014) and Chan et al. (2012), which were explained before in sections 3.3.2 and 3.3.3, respectively. The comparison focused on one step ahead forecasts through the root mean square error (RMSE) using the posterior mean as the estimate calculated on the second half of the data

$$RMSE = \sqrt{\frac{1}{n-s}\sum_{t=s+1}^{n}(y_t - \mathbb{E}(y_t|y_{1:t-1}, \boldsymbol{X}))^2},$$

where $\boldsymbol{X}$ is the design matrix and $s = \lfloor n/2 \rfloor$ is largest integer less than or equal to $n/2$.

3.3.6 Sparse autoregressive process for dynamic variable selection

Lastly, we discuss the approach of Ročková (2016) which was called *Autoregressive Spike-and-Slab Process* (ASSP) priors. Assuming the dynamic regression model with observation equation given by (3.29) with $\sigma_t^2 = 1$ and given a binary indicator $J_{j,t} \in \{0, 1\}$ and a lagged coefficient value $\beta_{j,t}$ the ASSP is defined as

$$\begin{aligned} p(\beta_{j,t}|J_{j,t}, \beta_{j,t-1}) &= (1 - J_{j,t})\psi_0(\beta_{j,t}|\lambda_0) + J_{j,t}\psi_0(\beta_{j,t}|\mu_{j,t}, \lambda_1), \\ \mu_{j,t} &= \phi_0 + \phi_1(\beta_{t-1} - \phi_0), \\ \theta_{j,t} &= p(J_{j,t} = 1|\beta_{j,t-1}), \end{aligned}$$

for $j = 1, ..., q$, where $|\phi_1| < 1$, $\psi_0(\beta_{j,t}|\lambda_0)$ is the spike distribution and $\psi_1(\beta_{j,t}|\mu_{j,t}, \lambda_1)$. It is assumed that the q time series $\beta_{1,1:n}, ..., \beta_{q,1:n}$ follow independent and identical ASSP priors.

The spike distribution can be any density sufficiently peaked around zero, for instance, the Laplace density $\psi_0(\beta_{j,t}|\lambda_0) = \frac{\lambda_0}{2}\exp -|\beta_{j,t}|\lambda_0$. The slab distribution $\psi_1(\beta_{j,t}|\mu_{j,t}, \lambda_1)$ should be moderately peaked around its mean $\mu_{j,t}$, where the amount of spread is regulated by $\lambda_1 > 0$. More specifically, the Gaussian density was chosen in the work, although other reasonable choices may be considered. It is worth to notice that the spike distribution does not depend on $\beta_{j,t-1}$. Under the spike, the coefficient time series is trivially stationarity, iid with a marginal density $\psi_0(\beta_{j,t}|\lambda_0)$, whereas under the slab it follows a stationarity Gaussian AR(1) process

$$\beta_{j,t} = \phi_0 + \phi_1(\beta_{j,t-1} - \phi_0) + e_t, \qquad e_t \sim \mathcal{N}(0, \lambda_1),$$

with stationary distribution

$$\psi_1^{ST} = \mathcal{N}\left(\phi_0, \frac{\lambda_1}{1-\phi^2}\right).$$

Although the process $\beta_{j,1:n}$ will be stationarity 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. That is why the sequence of mixture weights $\theta_{j,t}$ should move smoothly over time, depending not only on the previous value $\theta_{j,t-1}$ but also on $\beta_{j,t-1}$. It was assumed the following deterministic transition function for $\theta_{j,t}$, which preserves the spike-and-slab mixture for the marginals of $\beta_{j,t}$

$$\theta_{j,t} \equiv \theta(\beta_{j,t-1}) = \frac{\Theta\psi_1^{ST}(\beta_{j,t-1}|\lambda-1,\phi_0,\phi_1)}{\Theta\psi_1^{ST}(\beta_{j,t-1}|\lambda-1,\phi_0,\phi_1)+(1-\Theta)\psi_0(\beta_{j,t-1}|\lambda_0)},$$

where Θ is a marginal importance weight $0 < \Theta < 1$ which controls the overall balance between the spike and the slab distributions. The interpretation of Equation (3.3.6) is the following: if $|\beta_{j,t-1}|$ was large, then $\theta(\beta_{j,t-1})$ will be large as well, signaling that the current $\beta_{j,t-1}$ is more likely to be on the slab. The contrary occurs when $|\beta_{j,t-1}|$ is small.

It can be proven that (see Theorem 1 of Ročková (2016)) $\beta_{j,1:n}$ has a stationarity distribution

$$p^{ST}(\beta_{j,t}) = \Theta\psi_1^{ST}(\beta_{j,t}|\lambda-1,\phi_0,\phi_1)+(1-\Theta)\psi_0(\beta_{j,t}|\lambda_0),$$

that is, a spike-and-slab mixture model as noted before.

The practical appeal of the approach is the strategy pursued by the author, given the availability of MAP estimation algorithms. A one-step-late EM algorithm for MAP smoothing was implemented in a new dynamic coordinate-wise strategy. Although the algorithm may be fast, it gives only MAP estimates, nevertheless, the full distribution remains unavailable.

Chapter 4

Sparse covariance modeling

In this chapter we discuss some existing methods for sparse covariance modeling based on the regularization of the linear regressions which result from the covariance matrix decompositions. The objective is to briefly discuss the various decompositions of the covariance matrix that make the problem of estimation of a matrix into a linear regression problem, especially in the case of high-dimensional problems. In particular, we choose the modified Cholesky decomposition for the applications, because of its natural interpretation and practical appeal.

In Section 4.1, we present the modified Cholesky and other covariance matrix decompositions. Then, Section 4.2 is dedicated to presenting existing frequentist methods for regularizing the Cholesky factor such as the approach of Huang et al. (2006), based on the lasso regularization, the *Adaptative Banding with a nested lasso penalty* (AB) of Levina et al. (2008) and the *Forward Adaptative Banding* of Leng and Li (2011).

Lastly, in Section 4.3 some simulated and real data examples are presented comparing the Bayesian regularization of the Cholesky linear regressions based on the Normal-Gamma prior with the three methods introduced before.

4.1 Covariance decompositions

Let $\boldsymbol{y} = (\boldsymbol{y}_1, ..., \boldsymbol{y}_p)'$ a vector of p-dimensional responses, each one with n observations, with mean $\mathbf{0}$, and variance $\boldsymbol{\Sigma}$. The estimation of an unconditional covariance matrix (without constraints or structure) requires the estimation of $p(p-1)/2$ parameters. In many cases, the sample size n is not enough to estimate many parameters, especially if $p \gg n$. In this sense, there are some decompositions of $\boldsymbol{\Sigma}$ that allied to the use of linear models and regularization techniques such as the lasso, may reduce the number of parameters to be estimated.

The first of these is the variance-correlation decomposition

$$\boldsymbol{\Sigma} = \boldsymbol{D}\boldsymbol{R}\boldsymbol{D}, \tag{4.1}$$

where $\boldsymbol{D}$ is the diagonal matrix of standard deviations and $\boldsymbol{R}$ is the correlation matrix of $\boldsymbol{y}$. While the logarithm of $\boldsymbol{D}$ has no restrictions, $\boldsymbol{R}$ has the constraint of being positive definite with 1s on the main diagonal. The decomposition allows a separate estimation of $\boldsymbol{D}$ and $\boldsymbol{R}$, but it is inconvenient due to the restriction in $\boldsymbol{R}$.

Another possible decomposition of the covariance matrix is the object of study of the Gaussian graph models (Whittaker (2009))

$$\boldsymbol{\Sigma}^{-1} = \tilde{\boldsymbol{D}}\tilde{\boldsymbol{R}}\tilde{\boldsymbol{D}}, \tag{4.2}$$

where $\boldsymbol{\Sigma}^{-1}$ is the precision matrix, $\tilde{\boldsymbol{D}}$ is the diagonal matrix of the partial standard deviations and $\tilde{\boldsymbol{R}}$ is the partial correlation matrix of $\boldsymbol{y}$.

The precision matrix denotes the conditional dependencies. Assuming that $\boldsymbol{y}$ has mean zero, if the entry of $\boldsymbol{\Sigma}^{-1}$ is $\sigma_{ij}^{-1} = 0$, then the variables y_i and y_j are conditionally independent, given the other variables. This is shown in graph form, with the variables being equivalent to nodes, and the absence of arrows indicating conditional independence.

A third way of decomposing covariance matrices is through the very well known factor analysis. A factorial model for $\boldsymbol{y}$ with $E(\boldsymbol{y}) = \boldsymbol{\mu}$ and $\text{Var}(\boldsymbol{y}) = \boldsymbol{\Sigma}$, is given by

$$\boldsymbol{y} - \boldsymbol{\mu} = \boldsymbol{L}\boldsymbol{F} + \boldsymbol{\varepsilon}, \tag{4.3}$$

where $\boldsymbol{L} = (l_{ij})$ is a matrix $p \times q$ of coefficients of the q latent factors $\boldsymbol{F} = (f_1, ..., f_q)'$, and $\boldsymbol{\varepsilon} = (\epsilon_1, ..., \epsilon_p)'$ are the idiosyncratic errors. Thus, we can decompose $\boldsymbol{\Sigma}$ into

$$\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{\Lambda}\boldsymbol{L}' + \boldsymbol{\Psi}, \tag{4.4}$$

being $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ diagonal arrays of dimensions $q \times q$ and $p \times p$, respectively. Note that the problem of estimating $\boldsymbol{\Sigma}$ now turns out to estimating $(\boldsymbol{L}, \boldsymbol{\Lambda}, \boldsymbol{\Psi})$, and if q is relatively small in relation to p, there is a considerable reduction in the number of parameters to be estimated.

The Cholesky decompostion. Assume that $\boldsymbol{y}$ is an ordered random vector with mean zero and a positive definite covariance matrix $\boldsymbol{\Sigma}$. The standard Cholesky decomposition of $\boldsymbol{\Sigma}$ is as follows

$$\boldsymbol{\Sigma} = \boldsymbol{C}\boldsymbol{C}', \tag{4.5}$$

where $\boldsymbol{C} = (c_{ij})$ is a unique lower triangular matrix with positive diagonal entries.

The entries of $\boldsymbol{C}$ lacks of statistical interpretation. However, this can be solved by multiplying $\boldsymbol{C}$ by the inverse of $\boldsymbol{D}_1 = \text{diag}(c_{11}, ..., c_{pp})$. Using matrix multiplication, (4.5) can be rewritten as

$$\boldsymbol{\Sigma} = \boldsymbol{C}\boldsymbol{D}_1^{-1}\boldsymbol{D}_1^2\boldsymbol{D}_1^{-1}\boldsymbol{C}' = \boldsymbol{L}\boldsymbol{D}_1^2\boldsymbol{L}', \tag{4.6}$$

where $\boldsymbol{L} = \boldsymbol{C}\boldsymbol{D}_1^{-1}$ is obtained by dividing the jth column of $\boldsymbol{C}$ by the diagonal entry c_{jj}. The decomposition can also be stated as follows, known as *modified Cholesky decomposition* (see, e.g., Pourahmadi (2013)):

$$\boldsymbol{D} = \boldsymbol{T}\boldsymbol{\Sigma}\boldsymbol{T}', \qquad \boldsymbol{\Sigma}^{-1} = \boldsymbol{T}'\boldsymbol{D}^{-1}\boldsymbol{T}, \tag{4.7}$$

where $\boldsymbol{D} = \boldsymbol{D}_1^2$ and $\boldsymbol{T} = \boldsymbol{L}^{-1}$ is a lower triangular matrix with 1s as diagonal entries.

It is worth to note that Equation (4.4) is equivalent to (4.6), where the matrix $\boldsymbol{L}$ from Equation (4.4) is a lower triangular matrix $(p \times p)$ with 1s on the diagonal, $\boldsymbol{\Lambda} = \boldsymbol{D}_1^2$ and $\boldsymbol{\Psi} = \boldsymbol{0}$.

That is, the modified Cholesky decomposition is a particular case of a factorial model.

Lastly, we should note that the modified Cholesky decomposition transforms a restricted problem (the constraint of definite positiveness of $\boldsymbol{\Sigma}$) on an unrestricted regression problem, since the $\boldsymbol{L}$ and $\boldsymbol{\Lambda}$ has no restrictions.

To show that (4.7) has statistical interpretation, suppose that $\boldsymbol{y}$ is a time ordered random vector and let $\hat{y}_t$ be the linear least squares predictor of y_t based on its predecessors $y_{t-1}, ..., y_1$ and $\varepsilon_t = y_t - \hat{y}_t$ be its prediction error with variance $\sigma_t^2 = \text{Var}(\varepsilon_t)$. Then there are unique scalars ϕ_{tj} so that

$$y_t = \sum_{j=1}^{t-1} \phi_{tj} y_t + \varepsilon_t, \; t = 2, ...p, \tag{4.8}$$

where $y_1 = \varepsilon_1$.

Note that $\boldsymbol{y}$ could be sorted by any natural order such as time ordering and that it is clear that the matrices $\boldsymbol{T}$ (aka Cholesky factor) and $\boldsymbol{D}$ depend on the order of the p variables in $\boldsymbol{y}$.

Defining $\boldsymbol{\Phi}$ as

$$\boldsymbol{\Phi} = \begin{bmatrix} 0 & & & & \\ \phi_{21} & 0 & & & \\ \phi_{31} & \phi_{32} & 0 & & \\ ... & ... & ... & 0 & \\ \phi_{p1} & ... & ... & \phi_{p(p-1)} & 0 \end{bmatrix}, \tag{4.9}$$

one arrives at the decomposition (4.7), where $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_p)' = \boldsymbol{y} - \boldsymbol{\Phi y}$ is the vector of successive uncorrelated prediction errors such as $\boldsymbol{D} = \text{diag}(\sigma_1^2, ..., \sigma_p^2) = (\boldsymbol{I} - \boldsymbol{\Phi})\boldsymbol{\Sigma}(\boldsymbol{I} - \boldsymbol{\Phi})'$. That is, $\boldsymbol{T} = (\boldsymbol{I} - \boldsymbol{\Phi})$.

4.2 Sparse covariance modeling using Cholesky decomposition

An essential question is how to estimate $\boldsymbol{T}$ and $\boldsymbol{D}$ from the data. Estimates of the coefficients ϕ_{jl} can be computed easily for $p \leq n$. However, in large problems, it is expected that the estimates will be better by regularizing the Cholesky factor $\boldsymbol{T}$. For $p \gg n$, the regression problem is nonsingular in such a way that it becomes necessary to use shrinkage methods, which induce sparsity in $\boldsymbol{T}$.

The regularization of the Cholesky factor has been the subject of frequent study in the statistical literature and several proposals have already been made. Wu and Pourahmadi (2003) used nonparametric methods such as local polynomials to smooth the first k subdiagonals of $\boldsymbol{T}$, making the others equal zero. The proposed estimator was therefore based on a smoothed and banded matrix $\boldsymbol{T}$ and the number k was chosen by AIC criterion, considering a penalized Gaussian log likelihood.

Although the modified Cholesky decomposition does not require the response to be Normal, most authors assume data Gaussianity. In the article of Huang et al. (2006), the authors assume that $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$. The proposed method consists in minimizing the log likelihood (multiplied by -2), using a penalty based on the standard ℓ_1 norm, that is, the lasso method. The log-likelihood, considering the ordered data vector $\boldsymbol{y} = (y_1, ..., y_p)'$ is given by

$$-2\ell(\boldsymbol{\Sigma}, \boldsymbol{y}) = \log|\boldsymbol{\Sigma}| + \boldsymbol{y}'\boldsymbol{\Sigma}^{-1}\boldsymbol{y}. \tag{4.10}$$

Since $|\boldsymbol{\Sigma}| = |\boldsymbol{D}| = \prod_{j=1}^{p} \log \sigma_j^2$ and from (4.7), we have $\boldsymbol{\Sigma}^{-1} = \boldsymbol{T}'\boldsymbol{D}^{-1}\boldsymbol{T}$, (4.10) becomes

$$-2\ell(\boldsymbol{\Sigma}, \boldsymbol{y}) = \sum_{j=1}^{p} \log \sigma_j^2 + \sum_{j=1}^{p} \frac{\boldsymbol{\varepsilon}_j^2}{\sigma_j^2}.$$

Considering that $y_j = (y_{j1}, ..., y_{jn})'$, $j = 1, ..., p$, it follows that

$$-2\ell(\boldsymbol{\Sigma}, \boldsymbol{y}) = \sum_{j=1}^{p} \left(n \log \sigma_j^2 + \sum_{i=1}^{n} \frac{\varepsilon_{ij}^2}{\sigma_j^2} \right),$$

where $\varepsilon_{i1} = y_{i1}$ and $\varepsilon_{ij} = y_{ij} - \sum_{l=1}^{j-1} \phi_{jl} y_{il}$, for $j = 2, ..., p$. For a given $\lambda > 0$, the penalized log-likelihood, using the standard ℓ_1 norm is then

$$\sum_{j=1}^{p} \left(n \log \sigma_j^2 + \sum_{i=1}^{n} \frac{\varepsilon_{ij}^2}{\sigma_j^2} \right) + \lambda \sum_{j=2}^{p} \sum_{l=1}^{j-1} |\phi_{jl}|. \tag{4.11}$$

Note that minimizing (4.11) is equivalent to minimizing

$$\ell_j(\sigma_j, \boldsymbol{\phi}_j, y_{1j}, ..., y_{nj}) + \lambda P(\boldsymbol{\phi}_j) \tag{4.12}$$

separately, for $j = 1, ..., p$, where

$$\ell_j(\sigma_j, \boldsymbol{\phi}_j, y_{1j}, ..., y_{nj}) = \left(n \log \sigma_j^2 + \sum_{i=1}^{n} \frac{\varepsilon_{ij}^2}{\sigma_j^2} \right),$$

and $P(\boldsymbol{\phi}_j) = \sum_{l=1}^{j-1} |\phi_{jl}|$, $\boldsymbol{\phi}_j = (\phi_1, ..., \phi_{j,j-1})$, with $P(\phi_1) = 0$.

The Huang et al. (2006) approach therefore introduces arbitrarily localized zeros or at least generates shrinkage in the $\boldsymbol{T}$ elements, being more flexible than the Wu and Pourahmadi (2003) method that introduces bands, zeroing some more "distant" subdiagonals. However, as Levina et al. (2008) stands out, not necessarily the method proposed by them results in zeros in the precision matrix $\boldsymbol{\Sigma}^{-1}$, that is, the sparsity would be lost.

Considering this limitation of the method of Huang et al. (2006), Levina et al. (2008) introduced a new method of regularization of $\boldsymbol{T}$, which they called *Adaptive banding with a nested Lasso penalty*. In the proposed method, each equation has a different order, that is, considering a natural ordering of the variables, each variable y_j is regressed as a function of the closest predecessors $y_{j-k}, y_{j-k+1}, ..., y_{j-1}$, where $k = k_j$ is an order that varies between each equation j. The method preserves sparsity in $\boldsymbol{\Sigma}^{-1}$, since some successive elements of the subdiagonals of $\boldsymbol{T}$ are also zeroed, characterizing a banding type estimator.

The authors proposed replacing $\lambda P(\boldsymbol{\phi}_j)$ in (4.12) by the penalty

$$J_0(\phi_j) = \lambda \left(|\phi_{j,j-1}| + \frac{|\phi_{j,j-2}|}{|\phi_{j,j-1}|} + \frac{|\phi_{j,j-3}|}{|\phi_{j,j-2}|} + ... + \frac{|\phi_{j,1}|}{|\phi_{j,2}|} \right), \tag{4.13}$$

considering $0/0 = 0$. The effect of the penalty is that if the l-th variable is not included in the j-th equation ($\phi_{jl} = 0$), all subsequent variables ($l-1, l-2, ..., 1$) will be excluded from the regression, resulting in different orders $k = k_j$. Since the penalty involves a ratio between coefficients $\phi_{j,t}/\phi_{j,t-1}$, the measurement scale of variables can be a problem. However, for variables that are measured considering the same scale there would be no problem, otherwise a rescaling of the variables is required or even using different λ penalty parameters. For this reason, the authors proposed two modifications in the penalty (4.13), they are

$$J_1(\phi_j) = \lambda \left(\frac{|\phi_{j,j-1}|}{|\phi^*_{j,j-1}|} + \frac{|\phi_{j,j-2}|}{|\phi_{j,j-1}|} + \frac{|\phi_{j,j-3}|}{|\phi_{j,j-2}|} + ... + \frac{|\phi_{j,1}|}{|\phi_{j,2}|} \right) \tag{4.14}$$

$$J_2(\phi_j) = \lambda_1 \sum_{t=1}^{j-1} |\phi_{t,j}| + \lambda_2 \sum_{t=1}^{j-2} \frac{|\phi_{j,t}|}{|\phi_{j,t+1}|}, \tag{4.15}$$

where $\phi^*_{j,j-1}$ is the coefficient from regressing y_j on y_{j-1} alone.

It is worth to note that Huang et al. (2006) and Levina et al. (2008) have used the *K-fold* cross-validation method to choose λ, separating training and validation bases.

Motivated by the nested penalty of Levina et al. (2008), Leng and Li (2011) proposed a new approach for estimating $\boldsymbol{T}$ also using different orders k_j of the autoregressive Cholesky equations obtaining a banded Cholesky factor $\boldsymbol{T}$. The orders are obtained by minimizing a modified version of the BIC criterion

$$BIC = n \log |\hat{\boldsymbol{\Sigma}}| + \sum_{i=1}^{n} \boldsymbol{y}_i'(I - \hat{\boldsymbol{\Phi}})\hat{\boldsymbol{D}}^{-1}(I - \hat{\boldsymbol{\Phi}})\boldsymbol{y}_i + C_n \log{(n)} \sum_{j=1}^{p} k_j, \tag{4.16}$$

for $k_j \leq min\{n/(\log n)^2, j-1\}$, $j = 1, ..., p$, and a divergent sequence C_n. Note that $\boldsymbol{T}$ and $\boldsymbol{D}$ are obtained from the coefficients and residuals of the estimation of successive models AR (k_j).

The sum in (4.16) can be developed resulting in

$$BIC = \sum_{j=1}^{p} \left(n \log \hat{\sigma}_j^2 + \sum_{i=1}^{n} \frac{\hat{\epsilon}_{ij}^2}{\hat{\sigma}_j^2} + C_n \log{(n)} k_j \right), \tag{4.17}$$

where $\hat{\epsilon}_{ij} = \sum_{l=1}^{k_j} \phi_{j,j-l} y_{i,j-l}$. The equation (4.17) suggests that the minimization can be done separately for each subscript j.

The article shows that the constant C_n is chosen to accommodate large dimensions, as p increases, in addition to establishing some asymptotic properties. In their simulations, the authors used $C_n = max\{\log{(\log p)}, 1\}$, besides restricting the orders k_j to $[j/2]$, where $[a]$, the integer part of a.

Lastly, we should note that the variable selection and regularization methods methods presented in Chapter 2 can also be applied to estimate the coefficients from matrix $\boldsymbol{\Phi}$ or the factor $\boldsymbol{T}$ in each Cholesky equation $j = 2, ..., p$. In the next section we present the results of using the Normal Gamma prior, which is more general than the Bayesian lasso, to regularize these linear equations. We will show this in a real data application.

4.3 Simulated and real data examples

4.3.1 Simulations

In the same way as Leng and Li (2011), 50 replications of the multivariate normal $\mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma}_i)$ were performed, with the sample size of $n = 100$, considering three possible structures ($i = 1, 2, 3$) for the Cholesky factor $\boldsymbol{T}$ and a single structure for $\boldsymbol{D} = \text{diag}(0.01, ..., 0.01)$. In addition, we considered three numbers of variables $p = 10, 100, 200$. The first structure consists of a non-sparse $\boldsymbol{T}$ matrix, but with equal values in the sub-diagonals, which decreases as it moves away from the main diagonal (composed of 1s):

$$\mathbf{\Sigma}_1 : \phi_{j,j'} = 0.5^{|j-j'|}, j < j'.$$

The second structure consists of a sparse $\boldsymbol{T}$ matrix, containing only one sub-diagonal as follows

$$\begin{aligned}\mathbf{\Sigma}_2 : &\phi_{j,j-1} = 0.8,\\ &\phi_{j,j'} = 0,\ j' < (j-1).\end{aligned}$$

The third structure consists of a structure in which there are several orders k_j, one for each of the Cholesky sucessive equations, but with zeros arbitrarily located as

$$\begin{aligned}\mathbf{\Sigma}_3 : &k_j = \min\{\lfloor 6U \rfloor, j-1\},\\ &U \sim \mathcal{U}(0,1),\\ &Z \sim Bin(0.8),\end{aligned}$$

with

$$\begin{aligned}\phi_{j,j'} &= 0.5Z,\ (j-k_j) \leq j' \leq (j-1),\\ \phi_{j,j'} &= 0,\ (j-k_j) > j',\end{aligned}$$

where $\lfloor a \rfloor$ denotes the largest integer less than a, $\mathcal{U}$ denotes the Uniform distribution and Bin is the Binomial distribution.

Then, three methods explained in previous Section 4.2 were applied to the simulated data. As already mentioned, the penalty parameters λ of the Huang et al. (2006) and Levina et al. (2008) methods were obtained by the *K-fold* cross-validation method, with $K = 5$. Considering the Gaussianity assumption for the data, the parameters are obtained considering λ which minimizes

$$CV(\lambda) = \frac{1}{K}\sum_{v=1}^{K}\left(s_v \log|\hat{\mathbf{\Sigma}}_{-v}| + \sum_{i \in I_v} \boldsymbol{y}_i' \hat{\mathbf{\Sigma}}_{-v}^{-1} \boldsymbol{y}_i\right), \tag{4.18}$$

where I_v is the set of data subset indices S_v (validation data), s_v is the size of I_v and $\hat{\mathbf{\Sigma}}_v$ is the covariance matrix estimated using training data $(S - S_v)$, where S is the total data set. We considered $\lambda = 2^m$, $m = -2, -1.0, ..., 10$ and automatically chosen the penalty with the lowest cross validation error $CV(\lambda)$.

The results for the Kullback-Leiber loss measure $(L(\mathbf{\Sigma}, \hat{\mathbf{\Sigma}}))$ are shown in Table 4.1. The

Kullback-Leiber loss measure is calculated by

$$L(\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}}) = tr(\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}) - \log|\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}| - p.$$

Structure	p	Lasso		AB J1		FAB	
		mean	deviation	mean	deviation	mean	deviation
Σ_1	10	0.52	0.15	0.45	0.08	0.62	0.15
	100	8.83	0.89	6.03	0.45	8.49	0.60
	200	20.62	1.24	12.56	0.62	17.83	0.87
Σ_2	10	0.33	0.12	0.29	0.11	0.22	0.10
	100	6.72	0.59	2.55	0.29	2.24	0.26
	200	22.17	54.19	4.91	0.45	4.44	0.41
Σ_3	10	0.53	0.15	0.45	0.14	0.29	0.10
	100	17.94	1.48	13.06	3.23	7.15	0.50
	200	51.02	3.50	55.00	8.82	14.08	0.59

Table 4.1: *Kullback-Leiber results for Lasso, Adaptative Banding (AB J1) and Forward Adaptative Banding (FAB) - mean and standard deviation for 50 replications of the three covariance structures*

Note that the *Forward Adaptive Banding* (FAB) method, proposed by Leng and Li (2011), is superior to the other methods, except for the $\boldsymbol{\Sigma}_1$ matrix, which is not sparse. For this matrix the method of Levina et al. (2008) (AB, with penalty J1 as in (4.14)) is better. In addition, the FAB method has a much lower computational cost since there is no need for cross-validation. The lasso method of Huang et al. (2006), which introduces arbitrarily localized zeros, have the highest Kullback-Leiber among all and even for the $\boldsymbol{\Sigma}_3$ structure, whose simulation has zeros arbitrary located.

4.3.2 Empirical data application

It is well known that stock returns exhibit conditional heteroscedasticity. Thus, it makes sense to think of models with time-varying covariance matrices. In this section, the data of 92 daily log returns of shares that compose the S&P 100 index from 02/01/2004 to 10/28/2015 (n = 2977) will be described and the results of the Cholesky decomposition with regularization of the coefficient matrix $\boldsymbol{T}$ using the three methods described before as well as using the Normal-Gamma prior will be presented. Up till now, we assume that $\boldsymbol{D} = \boldsymbol{D}_t$ from decomposition (4.7) is time-varying, so regularization will be done in the standardized data $y_{ij}/\hat{\sigma}_{ij}$, where $\hat{\sigma}^2_{ij}$ is the estimate of stochastic volatility [1]. Figures 4.1 and 4.2 describe the raw and the standardized returns' data. We see that standardization corrected the conditional heteroscedasticity.

To understand if is reasonable to think that $\boldsymbol{T}$ is constant over time, but regularized, and $\boldsymbol{D} = \boldsymbol{D}_t$ is time-varying we randomly chose $m = 5$ companies from the 92 presented. For these companies, we applied the modified Cholesky decomposition to the raw data considering several time periods. The total period of 2977 days was divided into 99 periods of approximately 30 days, resulting in Figure 4.3. Note that while the coefficients $\phi_{j,l}$, $j = 2, ..., 5$; $l = 1, ..., j-1$ of the recursive regressions are more stable in time, the variances of errors terms are more volatile

[1] Gregor Kastner's *stochvol* package was used.

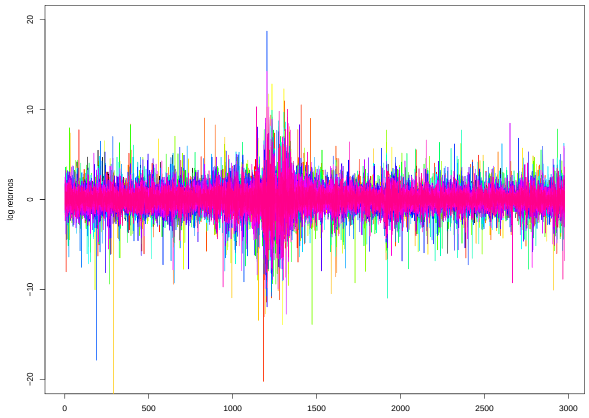


Figure 4.1: *Raw data of 92 log returns*

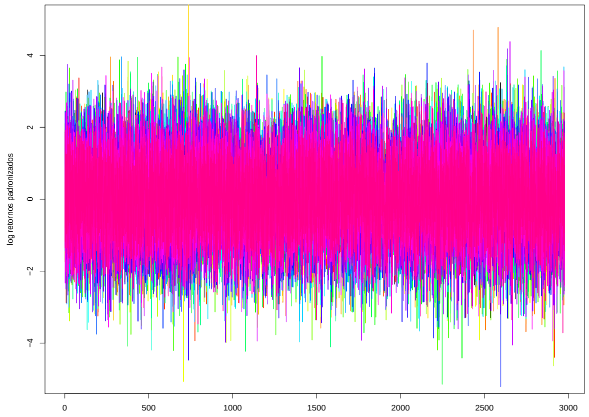


Figure 4.2: *Standardized data after estimating stochastic volatiliy*

than the coefficients.

Then, the same regularization methods previously discussed were applied to the standardized data. We also applied the Normal-Gamma prior, for which we have used the same prior specification discussed above in Section 4.3.1 and in Section 2.2.2 from Chapter 2.

The Normal-Gamma prior regularization was applied to the individual Cholesky equations stated by (4.8). That is, for each $t = 2, ..., p$ we assume that

$$y_t = \sum_{j=1}^{t-1} \phi_{tj} y_t + \varepsilon_t, \qquad \phi_{t,j} \sim \mathcal{N}(0, \psi_t), \qquad \psi_t \sim \mathcal{G}(\lambda, 1/2\gamma^2),$$

with $y_1 = \varepsilon_1$ and $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$.

We used the Gibbs sampler described in Algorithm 1 from Section 2.2.2 with 10000 iterations (and 5000 discarded as burnin), considering the following hyperparameter settings:

$$\sigma^2 \sim \mathcal{IG}(0.0001, 0.0001), \qquad \lambda \sim \mathcal{E}(1), \qquad \gamma^2|\lambda \sim \mathcal{IG}(2, M/2\lambda),$$

where M is chosen to be either equal to $M = \frac{1}{t-1}\sum_{j=1}^{t-1} \hat{\phi}_{tj}^2$, when the regression design matrix $\boldsymbol{X} = (y_1, ..., y_{t-1})$ is non singular, where $\hat{\phi}_{tj}$ is the least square estimate, or equal to the minimum length least squares estimate.

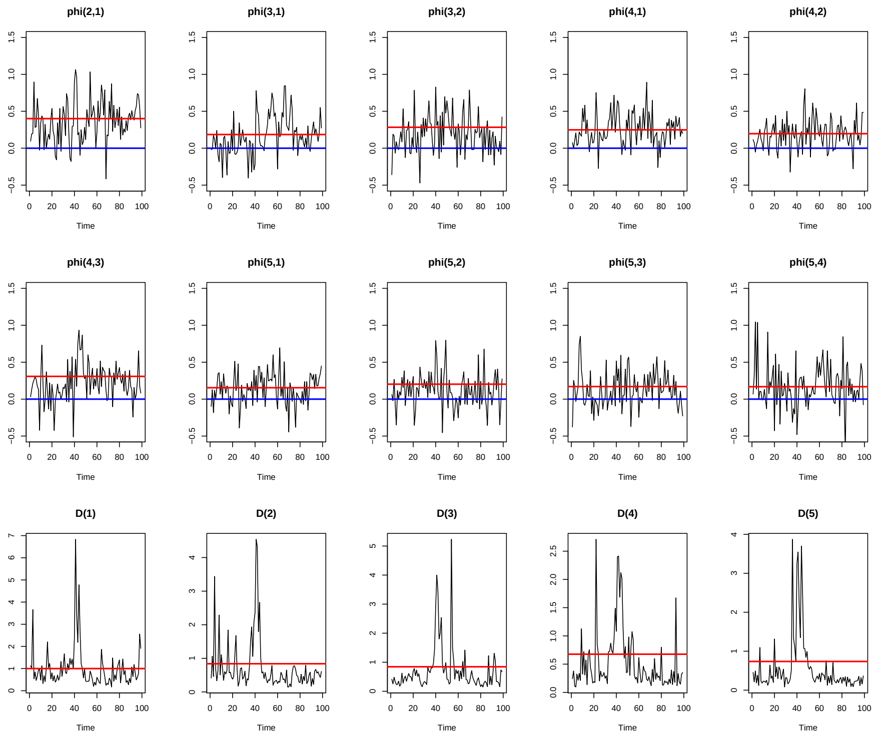
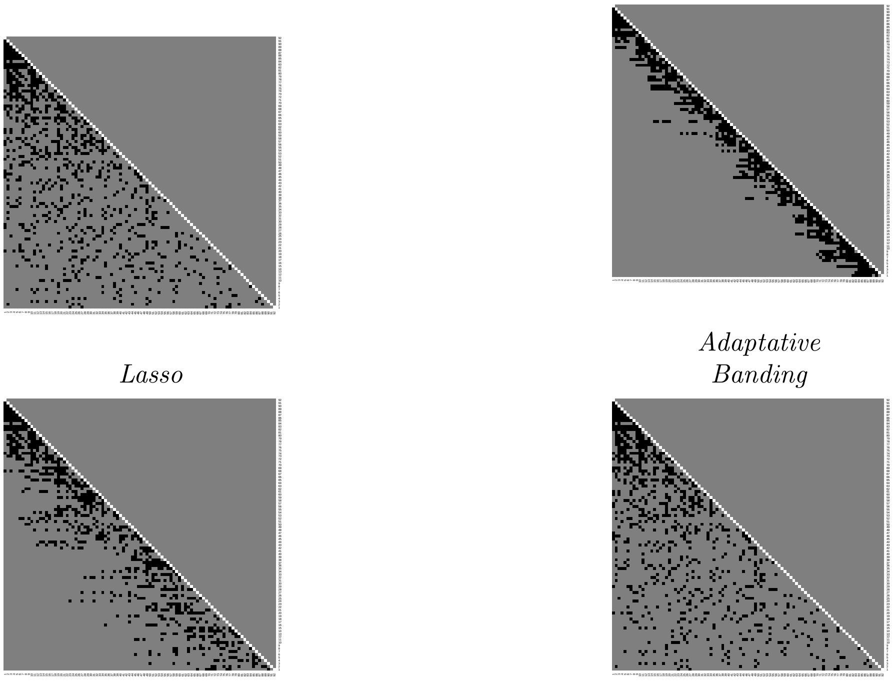


Figure 4.3: *Coefficients and variances that compose matrices* $\boldsymbol{T}$ *e* $\boldsymbol{D}$ *for the 5 randomly chosen returns*

The heatmaps of the Cholesky factor $\boldsymbol{T}$ are shown in Figure 4.4. The values marked in white, refer to the main diagonal, composed of 1s. The gray values refer to zeros. It is possible to notice that using the method *Adaptative banding* with the penalty J_1, the bands are more uniform than in the method *Forward adaptative banding*, for this set of data.

We also note that the Normal-Gamma prior heatmap is very similar to the Lasso of Huang et al. (2006). This is because the proposed prior method, as well the frequentist (or even Bayesian) Lasso is not a banding method. Of course, if we assume a natural order for the variables, it is worth to choose a banding method for estimating the covariance matrix, as the more closer are the variables the bigger will be the covariance between them. Otherwise, allowing for arbitrarily localized zeros in the Cholesky factor is appropriate in other situations.

This result, however, does not reveal much about the real fit to the data. Therefore, we calculated the sum of the squares of the residuals for each of the three methods, considering for each of the successive regressions, only the variables whose coefficients were not zeroed (considering a threshold of 0.0001, if the coefficient is less than this threshold then it is zero). The boxplots of the sums of the squares of the residuals are in Figure 4.5. It is noted that the fit of both the Lasso method and the NG prior seems more appropriate for this data.



Lasso

Adaptative Banding

Forward Banding

Normal-Gamma

Figure 4.4: *Heatmaps for Cholesky factor* $\boldsymbol{T}$

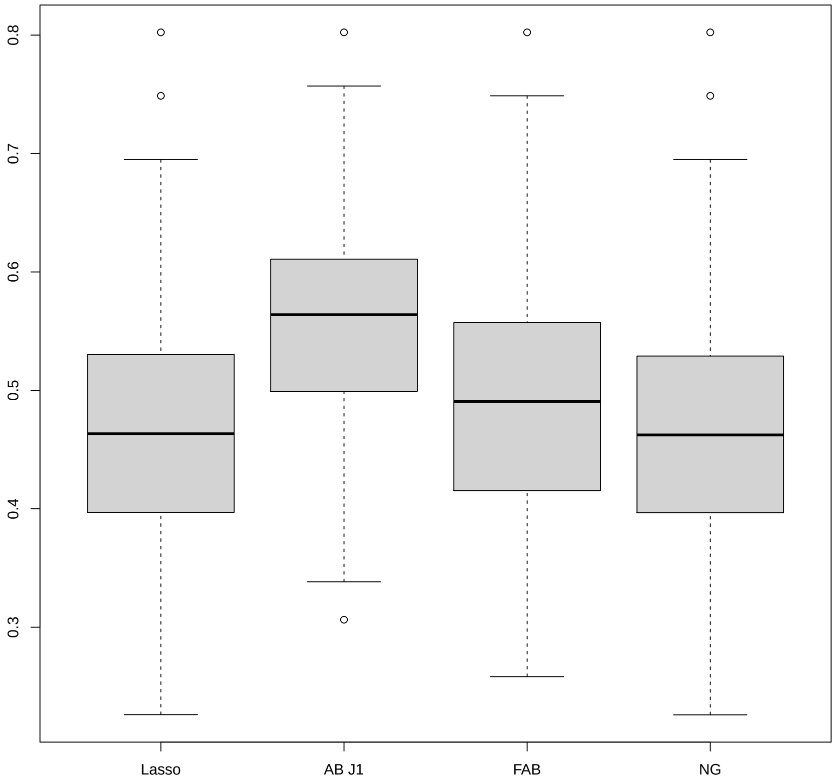


Figure 4.5: *Boxplot of the sum of square residuals of Lasso, Adaptative Banding (AB J1), Forward Adaptative Banding (FAB) and Normal-Gamma (NG) prior*

Chapter 5

A dynamic spike-and-slab model

In this chapter we propose a new method that accommodates time-varying sparsity, based on spike-and-slab priors. As noted in before, there are two types of sparsity in dynamic linear models: the *vertical sparsity*, which stands for time-varying subsets of relevant predictors, and the *horizontal sparsity*, which allows for intermittent zeros for when each individual predictor is not relevant at all times t. The proposed model aims to allow for vertical sparsity.

In Section 5.1, we discuss the basic formulation of the proposed model and its posterior inference. In Section 5.2, some simulated examples are given and the results of the dynamic spike-and-slab model is compared to some methods. Finally, we discuss empirical examples in Section 5.3.

5.1 A dynamic spike-and-slab model

We have seen in Section 3.3.2 that both references cited have adopted very similar strategies to induce sparsity in DLMs. Their basic approach was to rewrite the states' equation in terms of the scaled states $\tilde{\beta}_{j,t} = \beta_{j,t}/\omega_j$ thus arising at the non-centered parametrization, where the effect of the covariate $\boldsymbol{X}_j$ is divided into a fixed effect and a time-varying effect. Then, the shrinkage of the time-varying coefficients $\beta_{j,t}$ was done by assigning priors to their standard deviations ω_j. While Belmonte et al. (2014) used the Laplace prior for shrinking the standard deviations, Bitto and Frühwirth-Schnatter (2016) used the Normal-Gamma prior, which is more general since the Laplace prior is a special case from the Normal-Gamma prior where the shrinking parameter equals one.

Note that in both approaches the standard deviation ω_j plays the role of relevance of the jth predictor: small values of ω_j leads to greater shrinkage of the coefficient $\beta_{j,t}$ for all times t. That is, because the process standard deviation ω_j is taken as fixed for all times t, if it is pulled toward zero, then the (practically constant) effect of the covariate $\boldsymbol{X}_j$ is significant whenever the corresponding fixed regression effect is non-zero. In this sense, both approaches accommodates horizontal sparsity, as the shrinkage effect of the prior for ω_j is equal over all times t.

In other approaches discussed in Chapter 3 such as the *time-varying dimension* of Chan et al. (2012) and the *Normal-Gamma autoregressive process* of Kalli and Griffin (2014), the prior structures adopted accommodates vertical sparsity since the subset of covariates which were relevant at each time t was actually time-varying and modeled.

5.1.1 Model specification

In this section we explain and define our approach based on spike-and-slab priors. Our main contribution is developing a novel method that allows for vertical sparsity and it is applied to dynamic regression problems. In particular, we extend the previous work of Ishwaran and Rao (2005), which placed priors on the coefficients' variances. In this way, our prior can be view as a dynamic variable selection prior which induces either smoothness (through the slab) or shrinkage towards zero (through the spike) at each time point t.

Consider the Gaussian dynamic regression model with observation equation as

$$y_t = \boldsymbol{X}_t \boldsymbol{\beta}_t + \nu_t, \qquad \nu_t \sim \mathcal{N}(0, \sigma_t^2), \tag{5.1}$$

for $t = 1, ..., T$, where $\boldsymbol{X}_t$ is a $(T \times q)$ matrix of regressors, $\boldsymbol{\beta}_t$ is a $(q \times 1)$ vector of coefficients. We assume the following evolution equation for the scaled states $\tilde{\boldsymbol{\beta}}_{1:T} = (\tilde{\boldsymbol{\beta}}_1, ..., \tilde{\boldsymbol{\beta}}_T)$

$$\tilde{\boldsymbol{\beta}}_t = \boldsymbol{G}_t \tilde{\boldsymbol{\beta}}_{t-1} + \boldsymbol{\eta}_t, \qquad \boldsymbol{\eta}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{W}_t), \tag{5.2}$$

for $t = 2, ..., T$, where

$$\begin{aligned}
\tilde{\boldsymbol{\beta}}_t &= (\beta_{1,t}/\sqrt{\psi_{1,t}}, ..., \beta_{q,t}/\sqrt{\psi_{q,t}}),\\
\boldsymbol{G}_t &= \mathrm{diag}(\phi_1, ..., \phi_q),\\
\boldsymbol{W}_t &= \mathrm{diag}((1-\phi_1^2), ..., (1-\phi_q^2)),
\end{aligned}$$

where the initial condition for the scaled states is $\tilde{\boldsymbol{\beta}}_1 \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$.

The following definition specifies the generic dynamic spike-and-slab prior that can be assigned to the coefficients' variances $\boldsymbol{\psi}_{j,1:T}$ in order to induce shrinkage and/or variable selection.

Definition 5.1. ***Dynamic spike-and-slab prior.*** *Consider that $\psi_{j,t} = K_{j,t}\tau_j^2$. The dynamic spike-and-slab prior for $\beta_{j,1:T}$ is defined by (5.2) and by*

$$\begin{aligned}
\tau_j^2 &\overset{iid}{\sim} p(\tau_j^2|\boldsymbol{\theta}),\\
(K_{j,t}|K_{j,t-1} = \upsilon_i) &\overset{ind}{\sim} \omega_{j,1,i}\delta_{\upsilon_1}(.) + (1-\omega_{j,1,i})\delta_{\upsilon_0}(.),\\
\omega_{j,1,i} &= p\left(K_{j,t} = \upsilon_1 | K_{j,t-1} = \upsilon_i\right),
\end{aligned} \tag{5.3}$$

for $j = 1, ..., q$, $t = 2, ..., T$, where $\delta_x(.)$ is a discrete measure concentrated at value x, $\upsilon_i \in \{\upsilon_0, \upsilon_1\}$, $p(K_{j,1} = \upsilon_0) = p(K_{j,1} = \upsilon_1) = 1/2$ and $p(\tau_j^2|\boldsymbol{\theta})$ can be one of the mixing distributions discussed in Section 2.3.3.

As discussed before in Section 2.3.3, we assume that $\upsilon_0 = r$ (i.e., the ratio between the variances of the slab and the spike, which is a very small number less than 1) and that $\upsilon_1 = 1$. Furthermore, we assume a similar structure to that presented in Equation (2.42) from Section 2.3.3. That is, the vector of parameters $\boldsymbol{\theta}$ for τ_j^2 contains the hyperparameter Q_j which is distributed as

$$Q_j|\omega_{j,1,i} \overset{\text{ind}}{\sim} \mathcal{IG}(c_0, C_0/s^*(w)), \tag{5.4}$$

$j = 1, ..., q$, where $s^*(w) = s^*(w) = c[(1-\omega_{j,1,i})r + \omega_{j,1,i}]$ depends on the value of on the distribution constant c from Table 2.1 specified for $\tau_j^2|\boldsymbol{\theta}$ and on the value of $\omega_{j,1,i} = p\left(K_{j,t} = \upsilon_1 | K_{j,t-1} = \upsilon_i\right)$.

Because now we are talking about dynamic models and dynamic sparsity, we have a time-varying scale $\psi_{j,t}$, which is taken to be $\psi_{j,t} = K_{j,t}\tau_j^2$. That is, the time-varying pattern for the scale parameter is driven by the latent variable $K_{j,t}$, which evolves as a Markov switching process of order 1 and can assume two values $\upsilon_1 = 1$ or $\upsilon_0 = r$ accordingly to the transition matrix (which priorizes maintaining the same regime). For the other component τ_j^2 is placed a prior distribution (Inverse-Gamma, Gamma or Exponential) that together with the variable $K_{j,t}$ results in a spike-and-slab prior for $\psi_{j,t}$ that shrinks the coefficients $\beta_{j,t}$ whenever it gets a small value through the spike component of the mixture prior. Furthermore, by defining the hyperparameters c_0 and C_0 appropriately we can learn about Q_j and therefore about τ_j^2.

For instance, if $p(\tau_j^2|\boldsymbol{\theta}) \sim \mathcal{IG}(\nu, Q_j)$, then

$$\beta_{j,t}|\omega_{j,1,i} \sim \omega_{j,1,i}t_{2\nu}(0, Q_j/\nu) + (1-\omega_{j,1,i})t_{2\nu}(0, rQ_j/\nu), \tag{5.5}$$

and the constant c in (5.4) is equal $1/(\nu-1)$. We will refer to this as the *dynamic NMIG prior.*

If $p(\tau_j^2|\boldsymbol{\theta}) \sim \mathcal{G}(a, 1/2Q_j)$, then

$$\beta_{j,t}|\omega_{j,1,i} \sim \omega_{j,1,i}\mathcal{NG}(\beta_{j,t}|a, Q_j) + (1-\omega_{j,1,i})(\beta_{j,t}|a, r, Q_j), \tag{5.6}$$

and the constant c in (5.4) is equal $2a$. We will refer to this as the *dynamic NG prior.*

Finally, if we assume that $p(\tau_j^2|\boldsymbol{\theta}) \sim \mathcal{E}(1/2Q_j)$, then

$$\beta_{j,t}|\omega_{j,1,i} \sim \omega_{j,1,i}Lap(\sqrt{Q_j}) + (1-\omega_{j,1,i})Lap(\sqrt{rQ_j}), \tag{5.7}$$

and the constant c in (5.4) is equal 2. We will refer to this as the *dynamic Laplace prior.*

The assumption that Q_j is given a prior $Q_j \sim \mathcal{IG}(c_0, C_0/s^*(w))$ as defined in (5.4) makes τ_j^2 also indirectly depend on the previous value of the Markov latent variable $K_{j,t-1} = \upsilon_i$. The purpose of this is to keep the variance of the coefficients $\beta_{j,t}$ at each time point t constant across the priors defined by equations (5.5), (5.6) and (5.7) as discussed before in Section 2.3.3.

Stationary AR(1) process. It is worth noting that each process $\beta_{j,t}$ is independent of $\beta_{l,t}$ for any $l, j \in \{1, ..., q\}$ and that (5.2) can be written as

$$\beta_{j,t} = \sqrt{\frac{\psi_{j,t}}{\psi_{j,t-1}}}\phi_j\beta_{j,t-1} + \epsilon_j, \quad \epsilon_{j,t} \sim \mathcal{N}(0, \psi_{j,t}(1-\phi_j^2)), \tag{5.8}$$

for $t = 2, ..., T$ and $j = 1, ..., q$, with $\beta_{j,1} \sim \mathcal{N}(0, \psi_{j,1})$. That is, $\tilde{\beta}_{j,t} = \beta_{j,t}/\sqrt{\psi_{j,t}}$, conditional on $\boldsymbol{\psi}_{j,1:T} = (\psi_{j,1}, ..., \psi_{j,T})$ is

$$\tilde{\beta}_{j,t} = \phi_j\tilde{\beta}_{j,t-1} + \eta_{j,t}, \quad \eta_{j,t} \sim \mathcal{N}(0, (1-\phi_j^2)),$$

which is a stationary AR(1) process assuming that the autoregressive parameter $|\phi_j| < 1$. Thus, by the stationary assumption, the conditional (on $\psi_{j,t}$) variance of $\tilde{\beta}_{j,t}$ equals 1, since

$$\text{Var}\left(\frac{\beta_{j,t}}{\psi_{j,t}}\middle|\psi_{j,t}\right) = \text{Var}\left(\frac{\beta_{j,t-1}}{\psi_{j,t-1}}\middle|\psi_{j,t}\right),$$

which occurs if, and only if, each $\beta_{j,t}|\psi_{j,t} \sim \mathcal{N}(0, \psi_{j,t})$ with $\psi_{j,t}$ following the process defined by 5.1.

Markov latent variables. Note that (5.8) is inspired in the NGAR process from Kalli and Griffin (2014). Nevertheless, we assume a Markov switching structure for $\psi_{j,t}$ instead of the GAR process. That is, $K_{j,t}$ is a latent variable that can assume binary values (regimes) r or 1, depending only on the previous value of $K_{j,t-1} = \upsilon_i \in \{r, 1\}$, so that, defining the density $p(\tau_j^2|\boldsymbol{\theta})$ from Definition 5.1 and noting that $\psi_{j,t} = K_{j,t}\tau_j^2$, we have

$$(\psi_{j,t}|K_{j,t-1} = \upsilon_i) \sim \omega_{j,1,i} p_{slab}(\psi_j|Q_j) + (1 - \omega_{j,1,i}) p_{spike}(\psi_j|r, Q_j), \tag{5.9}$$

where $\omega_{j,1,i}$ is the transition probability of the first order Markov process $K_{j,t}$ to regime $\upsilon_1 = 1$ given that $K_{j,t-1} = \upsilon_i \in \{r, 1\}$. Thus, by adopting a regime switching model, the process $\psi_{j,t}$ can switch between the spike and the slab variances' distributions according to the following transition probabilities

$$\boldsymbol{\mathcal{P}}_j = \begin{bmatrix} \omega_{j,0,0} & \omega_{j,0,1} \\ \omega_{j,1,0} & \omega_{j,1,1} \end{bmatrix}$$

where $\omega_{j,k,i} = P(K_{j,t} = \upsilon_k|K_{j,t-1} = \upsilon_i)$ denotes the probability of $K_{j,t}$ changing to regime υ_k from regime υ_i, $k, i \in \{0, 1\}$. Note that $\omega_{j,0,1} = (1 - \omega_{j,1,1})$ and $\omega_{j,1,0} = (1 - \omega_{j,0,0})$.

State space representation. The Gaussian dynamic regression model with an appropriate shrinking prior for the variances $\boldsymbol{\psi}_{j,1:T} = (\psi_{j,1}, ..., \psi_{j,T})$, considering q potential predictors and T time points, depends on the following parameters

$$\boldsymbol{\Theta} = (\sigma_1^2, ..., \sigma_T^2, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_T, \boldsymbol{K}_1, ..., \boldsymbol{K}_T, \boldsymbol{\tau}^2, \boldsymbol{Q}, \boldsymbol{\mathcal{P}}), \tag{5.10}$$

where $\boldsymbol{\beta}_t = (\beta_{1,t}, ..., \beta_{q,t})'$, $\boldsymbol{K}_t = (K_{1,t}, ..., K_{q,t})'$, $\boldsymbol{\tau} = (\tau_1^2, ..., \tau_q^2)$, $\boldsymbol{Q} = (Q_1, ..., Q_q)$, and the collection of transition matrices $\boldsymbol{\mathcal{P}} = (\boldsymbol{\mathcal{P}}_1, ..., \boldsymbol{\mathcal{P}}_q)$.

For a simpler specification, from now on we assume that the observation variance is constant over time such that $\sigma_t^2 = \sigma^2$. Extending the model to accommodate stochastic volatility is straightforward. Thus, assuming constant observation variance, we have two sequences of time-varying parameters: the coefficients $\boldsymbol{\beta}_{1:T}$ and the binary variables $\boldsymbol{K}_{1:T}$, which refers to Section 3.3.4 where the dynamic mixture models of Gerlach et al. (2000) were discussed. Based on the Equation (3.27), the state space representation of the proposed model is

$$\begin{aligned} y_t &= \boldsymbol{F}_t'\tilde{\boldsymbol{\beta}}_t + \gamma_t u_t, & u_t &\sim \mathcal{N}(0, 1), \\ \tilde{\boldsymbol{\beta}}_t &= \boldsymbol{G}_t\tilde{\boldsymbol{\beta}}_{t-1} + \boldsymbol{\Gamma}_t\boldsymbol{v}_t, & \boldsymbol{v}_t &\sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}), \end{aligned} \tag{5.11}$$

with $\boldsymbol{F}_t' = (X_{1,t}\sqrt{\psi_{1,t}}, ..., X_{q,t}\sqrt{\psi_{q,t}})$, $\gamma_t = \sigma$, $\boldsymbol{\Gamma}_t = \boldsymbol{W}_t^{1/2} = \text{diag}\left(\sqrt{(1-\phi_1^2)}, ..., \sqrt{(1-\phi_q^2)}\right)$ and $\boldsymbol{G}_t = \text{diag}(\phi_1, ..., \phi_q)$, where $\boldsymbol{W}_t$, $\boldsymbol{G}_t$ and $\tilde{\boldsymbol{\beta}}_t$ are the same from (5.2). Thus, one can see that matrix $\boldsymbol{F}_t'$ depends on the values of $\boldsymbol{\psi}_t$ which in turn depend on the binary latent variables $\boldsymbol{K}_t$. Thus, we can use the sampling scheme presented in Algorithm 6 from Section 3.3.4 to draw $\boldsymbol{K}_{1:T}$ without conditioning on $\tilde{\boldsymbol{\beta}}_{1:T}$.

Other prior specifications. In order to complete the specification, we shall assign prior distributions to parameters σ^2, $\boldsymbol{\phi}$ and to the transition probabilities $\boldsymbol{\mathcal{P}}$ in a full Bayes strategy. For the observation variance σ^2, we assume the conjugate traditional prior

$$\sigma^2 \sim \mathcal{IG}(a_\sigma, b_\sigma). \tag{5.12}$$

For the AR parameters $\boldsymbol{\phi}$, we assume that each ϕ_j are independent from each other and distributed as

$$\phi_j \sim \mathcal{B}(a_\phi, b_\phi), \tag{5.13}$$

for which we are not considering the case $-1 < \phi_j < 0$. Finally, for the transition probabilities $\mathcal{P}_j$ we also give independent Beta distributions as

$$\omega_{j,i,i} \sim \mathcal{B}(a_\omega, b_\omega), \tag{5.14}$$

for $j = 1, ..., q$, $i \in \{0, 1\}$, remembering that $\omega_{j,k,i} = (1 - \omega_{j,i,i})$, $k \neq i$, $k, i \in \{0, 1\}$.

The directed acyclic graph (DAG) that summarizes the dependencies of the proposed model is shown in Figure 5.1.

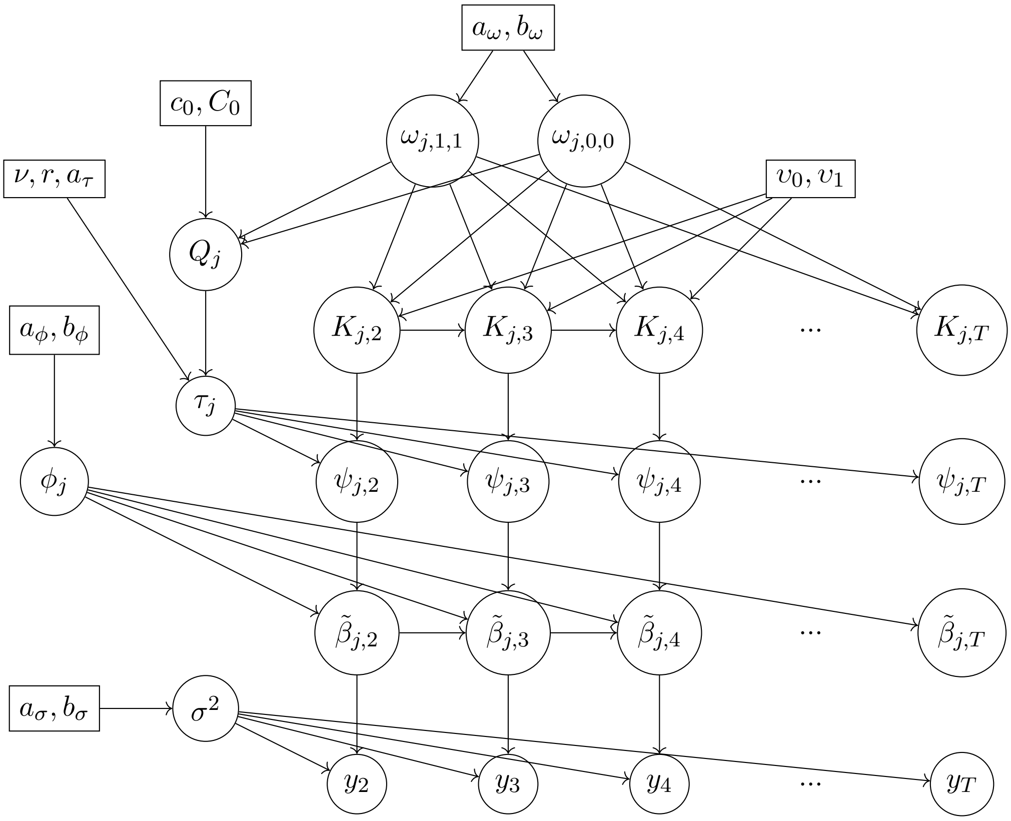


Figure 5.1: *Dependence structure for dynamic spike-and-slab model*

5.1.2 Posterior inference

The posterior distribution of the parameters can be drawn using an hybrid Gibbs sampler with an additional Metropolis-Hastings update. The scaled states $\tilde{\boldsymbol{\beta}}_1, ..., \tilde{\boldsymbol{\beta}}_T$ can be updated using the FFBS algorithm within the Gibbs sampler as presented in Algorithm 5 of Section 3.2.2, while the process $\boldsymbol{K} = (\boldsymbol{K}_1, ..., \boldsymbol{K}_T)$ is updated using the algorithm of Gerlach et al. (2000) showed in Algorithm 6 of Section 3.3.4. The full conditionals are given below.

First, for the parameters $\boldsymbol{\tau}^2$ and $\boldsymbol{Q}$ note that for each $j = 1, ..., q$ we have

$$\begin{aligned} p(\tau_j^2|\boldsymbol{\Theta}_{\backslash\tau_j^2}, \boldsymbol{y}) &\propto p(\tau_j^2|\boldsymbol{\theta})p(\boldsymbol{\beta}_j|\tau_j^2, \boldsymbol{K}_j, \phi_j) \\ &= p(\tau_j^2|\boldsymbol{\theta})p(\beta_{j,1}|K_{j,1}, \tau_j^2)\prod_{t=2}^{T} p(\beta_{j,t}|\beta_{j,t-1}, \tau_j^2, \boldsymbol{K}_j, \phi_j), \end{aligned}$$

where $\boldsymbol{\Theta}_{\backslash\tau_j^2}$ denotes all the parameters specified in (5.10) except from τ_j^2 and $\boldsymbol{y} = (y_1, ..., y_T)'$. The term $p(\tau_j^2|\boldsymbol{\theta})$ is the prior of τ_j^2, while the term $\prod_{t=2}^{T} p(\beta_{j,t}|\beta_{j,t-1}, \tau_j^2, \boldsymbol{K}_j, \phi_j)$ is the product of Normal densities

$$(\beta_{j,t}|\psi_{j,t}, \psi_{j,t-1}, \phi_j) \sim \mathcal{N}\left(\sqrt{\frac{\psi_{j,t}}{\psi_{j,t-1}}}\phi_j\beta_{j,t-1}, \psi_{j,t}(1-\phi_j^2)\right)$$

as in (5.8). Therefore, for the dynamic NMIG prior specified by (5.5) we have

$$\begin{aligned} p(\tau_j^2|\boldsymbol{\Theta}_{\backslash\tau_j^2}, \boldsymbol{y}) \propto\ & \frac{Q_j^{\nu}}{\Gamma(\nu)}\tau_j^{2(-\nu-1)}\exp\left\{-\frac{Q_j}{\tau_j^2}\right\}\psi_{j,1}^{-1/2}\exp\left\{-\frac{\beta_{j,1}^2}{2\psi_{j,1}(1-\phi_j^2)}\right\} \\ & \times \psi_{j,2}^{-1/2}\exp\left\{-\frac{\left(\beta_{j,2}-\sqrt{\frac{\psi_{j,2}}{\psi_{j,1}}}\phi_j\beta_{j,1}\right)^2}{2\psi_{j,2}(1-\phi_j^2)}\right\} \times (...) \\ & \times \psi_{j,T}^{-1/2}\exp\left\{-\frac{\left(\beta_{j,T}-\sqrt{\frac{\psi_{j,T}}{\psi_{j,T-1}}}\phi_j\beta_{j,T-1}\right)^2}{2\psi_{j,T}(1-\phi_j^2)}\right\} \\ =\ & \frac{Q_j^{\nu}}{\Gamma(\nu)}\tau_j^{2(-\nu-1)}\exp\left\{-\frac{Q_j}{\tau_j^2}\right\}(\psi_{j,1}\psi_{j,2}...\psi_{j,T})^{-1/2} \\ & \times \exp\left\{-\sum_{t=1}^{T}\frac{\left(\beta_{j,t}-\sqrt{\frac{\psi_{j,t}}{\psi_{j,t-1}}}\phi_j\beta_{j,t-1}\right)^2}{2\psi_{j,t}(1-\phi_j^2)}\right\}, \end{aligned}$$

where we assume that $\mathbb{E}(\beta_{j,1}) = 0$ as defined before. Because each $\psi_{j,t} = K_{j,t}\tau_j^2$, then

$$
\begin{aligned}
p(\tau_j^2|\boldsymbol{\Theta}_{\backslash\tau_j^2},\boldsymbol{y}) &\propto \tau_j^{2(-\nu-1)}\exp\left\{-\frac{Q_j}{\tau_j^2}\right\}(K_{j,1}\tau_j^2K_{j,2}\tau_j^2...,K_{j,T}\tau_j^2)^{-1/2}\\
&\times\exp\left\{-\frac{1}{2\tau_j^2}\sum_{t=1}^{T}\frac{\left(\beta_{j,t}-\sqrt{\dfrac{K_{j,t}}{K_{j,t-1}}}\phi_j\beta_{j,t-1}\right)^2}{K_{j,t}(1-\phi_j^2)}\right\}\\
&=\tau_j^{2(-\nu-T/2-1)}\exp\left\{-\frac{Q_j}{\tau_j^2}-\frac{1}{2\tau_j^2}\sum_{t=1}^{T}\frac{\left(\beta_{j,t}-\sqrt{\dfrac{K_{j,t}}{K_{j,t-1}}}\phi_j\beta_{j,t-1}\right)^2}{K_{j,t}(1-\phi_j^2)}\right\},
\end{aligned}
$$

where $\beta_{j,0}=\mathbb{E}(\beta_{j,1})=0$. That is, the full conditional of τ_j^2 is

$$
(\tau_j^2|\boldsymbol{\Theta}_{\backslash\tau_j^2},\boldsymbol{y})\sim\mathcal{IG}\left(\nu+\frac{T}{2},Q_j+\frac{1}{2}\sum_{t=1}^{T}\frac{\left(\beta_{j,t}-\sqrt{\dfrac{K_{j,t}}{K_{j,t-1}}}\phi_j\beta_{j,t-1}\right)^2}{K_{j,t}(1-\phi_j^2)}\right). \tag{5.15}
$$

If we assume the dynamic NG prior from (5.6), the prior for $\tau_j^2|\boldsymbol{\theta}\sim\mathcal{G}(a_\tau,1/2Q_j)$. Then,

$$
\begin{aligned}
p(\boldsymbol{\Theta}_{\backslash\tau_j^2},\boldsymbol{y}) &\propto \tau_j^{2(a_\tau-T/2-1)}\exp\left\{-\frac{\tau_j^2}{2Q_j}-\frac{1}{2\tau_j^2}\sum_{t=1}^{T}\frac{\left(\beta_{j,t}-\sqrt{\dfrac{K_{j,t}}{K_{j,t-1}}}\phi_j\beta_{j,t-1}\right)^2}{K_{j,t}(1-\phi_j^2)}\right\},\\
&\propto \tau_j^{2(p-1)}\exp\left\{-\frac{(g\tau_j^2+h/\tau_j^2)}{2}\right\},
\end{aligned}
$$

with

$$
g=1/Q_j,\quad h=\sum_{t=1}^{T}\frac{\left(\beta_{j,t}-\sqrt{\dfrac{K_{j,t}}{K_{j,t-1}}}\phi_j\beta_{j,t-1}\right)^2}{K_{j,t}(1-\phi_j^2)},\quad p=a_\tau-T/2.
$$

That is, the full conditional of τ_j^2 is

$$
(\tau_j^2|\boldsymbol{\Theta}_{\backslash\tau_j^2},\boldsymbol{y})\sim\mathcal{GIG}(p,g,h), \tag{5.16}
$$

where $\mathcal{GIG}(p,g,h)$ is the Generalized Inverse Gaussian distribution with probability density function

$$
p(x)=\frac{(g/h)^{p/2}}{2K_p(\sqrt{gh})}x^{(p-1)}\exp\left\{-\frac{(gx+h/x)}{2}\right\},\quad x>0,
$$

where K_p is a modified Bessel function of the second kind and $g > 0$, $h > 0$, $p \in \mathbb{R}$.

Lastly, for the dynamic Laplace mixture prior from Equation (5.7), as the Exponential distribution is equivalent to a Gamma distribution with shape parameter equal $a_\tau = 1$, the full conditional τ_j^2 is equal to

$$(\tau_j^2|\mathbf{\Theta}_{\setminus\tau_j^2}, \boldsymbol{y}) \sim \mathcal{GIG}\left(1 - T/2, 1/Q_j, \sum_{t=1}^{T} \frac{\left(\beta_{j,t} - \sqrt{\frac{K_{j,t}}{K_{j,t-1}}}\phi_j\beta_{j,t-1}\right)^2}{K_{j,t}(1-\phi_j^2)}\right). \tag{5.17}$$

For each parameter Q_j, note by (5.4) that is given a conditional prior (as $s^*(w)$ depends on the parameter $\omega_{j,1,i}$). Then, for the dynamic NMIG prior

$$\begin{aligned}
p(Q_j|\mathbf{\Theta}_{\setminus Q_j}, \boldsymbol{y}) &\propto p(Q_j|\omega_{j,1,i})p(\tau_j^2|Q_j) \\
&= \frac{[C_0/s^*(w)]^{c_0}}{\Gamma(c_0)} Q_j^{-c_0-1} \exp\left\{-\frac{[C_0/s^*(w)]}{Q_j}\right\} \frac{Q_j^\nu}{\Gamma(\nu)} (\tau_j^2)^{-\nu-1} \exp\left\{-\frac{Q_j}{\tau_j^2}\right\} \\
&\propto Q_j^{-c_0-1+\nu} \exp\left\{-\frac{2\tau_j^{-2}Q_j + 2[C_0/s^*(w)]/Q_j}{2}\right\},
\end{aligned}$$

which is a $\mathcal{GIG}(p, g, h)$ distribution with $p = \nu - c_0$, $g = 2\tau_j^{-2}$ and $h = 2[C_0/s^*(w)]$.

For the dynamic NG prior and the dynamic Laplace prior, we have

$$\begin{aligned}
p(Q_j|\mathbf{\Theta}_{\setminus Q_j}, \boldsymbol{y}) &\propto p(Q_j|\omega_{j,1,i})p(\tau_j^2|Q_j) \\
&\propto Q_j^{-c_0-1} \exp\left\{-\frac{[C_0/s^*(w)]}{Q_j}\right\} \frac{(1/2Q_j)^{a_\tau}}{\Gamma(a_\tau)} (\tau_j^2)^{a_\tau-1} \exp\left\{-\frac{\tau_j^2}{2Q_j}\right\} \\
&= Q_j^{-c_0-a_\tau-1} \exp\left\{-\frac{\tau_j^2/2 + [C_0/s^*(w)]}{Q_j}\right\},
\end{aligned}$$

which is a $\mathcal{IG}(c_0 + a_\tau, \tau_j^2/2 + [C_0/s^*(w)])$ distribution and where $a_\tau = 1$ for the Laplace prior.

For the observation variance σ^2, we have

$$\begin{aligned}
p(\sigma^2|\mathbf{\Theta}_{\setminus\sigma^2}) &\propto p(\sigma^2|a_\sigma, b_\sigma)p(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2) \\
&\propto \sigma^{2(-a_\sigma-1)} \exp\left(-\frac{b_\sigma}{\sigma^2}\right) \sigma^{2(-T/2)} \exp\left\{-\frac{1}{2}\sum_{t=1}^{T} \frac{(y_t - \boldsymbol{X}_t\boldsymbol{\beta}_t)^2}{\sigma^2}\right\},
\end{aligned}$$

which gives

$$(\sigma^2|\mathbf{\Theta}_{\setminus\sigma^2}, \boldsymbol{y}) \sim \mathcal{IG}\left(a_\sigma + \frac{T}{2}, b_\sigma + \frac{1}{2}\sum_{t=1}^{T}(y_t - \boldsymbol{X}_t\boldsymbol{\beta}_t)^2\right). \tag{5.18}$$

For the AR parameters $\boldsymbol{\phi}$, note that for each ϕ_j, $j = 1, ..., q$, we have

$$\begin{aligned} p(\phi_j|\boldsymbol{\Theta}_{\backslash\phi_j},\boldsymbol{y}) &\propto p(\phi_j|a_\phi,b_\phi)p(\boldsymbol{\beta}_j|\boldsymbol{K}_j,\sigma^2,\tau_j^2) \\ &\propto \phi_j^{(a_\phi-1)}(1-\phi_j)^{(b_\phi-1)}\exp\left\{-\sum_{t=1}^{T}\frac{\left(\beta_{j,t}-\sqrt{\frac{\psi_{j,t}}{\psi_{j,t-1}}}\phi_j\beta_{j,t-1}\right)^2}{2\psi_{j,t}(1-\phi_j^2)}\right\}, \end{aligned} \tag{5.19}$$

which is a non-standard form, thus, ϕ_j can be updated using a Metropolis-Hastings update as follows. As ϕ_j can not assume the values 0 and 1, we generate, for each MCMC iteration m, the candidate ϕ_j^* using the following Beta proposal density $q\left(\phi_j^*|\phi_j^{(m-1)}\right)$

$$\phi_j^* \sim \mathcal{B}\left(\alpha,\xi\left(\phi_j^{(m-1)}\right)\right), \quad \xi\left(\phi_j^{(m-1)}\right)=\alpha\left(\frac{1-\phi_j^{(m-1)}}{\phi_j^{(m-1)}}\right), \tag{5.20}$$

where α is a tuning parameter and $q\left(\phi_j^*|\phi_j^{(m-1)}\right)$ is the density function of the Beta distribution in (5.20). The acceptance distribution $\mathcal{A}\left(\phi_j^*|\phi_j^{(m-1)}\right)$ is

$$\mathcal{A}\left(\phi_j^*|\phi_j^{(m-1)}\right)=min\left\{1,\frac{f\left(\phi_j^*\right)q\left(\phi_j^{(m-1)}|\phi_j^*\right)}{f\left(\phi_j^{(m-1)}\right)q\left(\phi_j^*|\phi_j^{(m-1)}\right)}\right\},$$

with $f\left(\phi_j^*\right)$, $f\left({\phi_j}^{(m-1)}\right)$ obtained from the full conditional in (5.19), and $q\left(\phi_j^{(m-1)}|\phi_j^*\right)$, $q\left(\phi_j^*|\phi_j^{(m-1)}\right)$ obtained from the proposal density in (5.20). That is, with probability $\mathcal{A}\left(\phi_j^*|\phi_j^{(m-1)}\right)$ we set $\phi_j^{(m)}=\phi_j^*$, otherwise we set $\phi_j^{(m)}=\phi_j^{(m-1)}$.

For each transition probabilities $\omega_{j,1,1}$ and $\omega_{j,0,0}$, that compose $\boldsymbol{\mathcal{P}}_j$, the full conditionals are

$$\begin{aligned} p(\omega_{j,i,i}|\boldsymbol{\Theta}_{\backslash\omega_{j,i,i}},\boldsymbol{y}) &\propto p(\omega_{j,i,i}|a_\omega,b_\omega)p(\boldsymbol{K}_j|\boldsymbol{\mathcal{P}}_j) \\ &\propto {\omega_{j,i,i}}^{(a_\omega-1)}(1-\omega_{j,i,i})^{(b_\omega-1)}\prod_{t=2}^{T}p(K_{j,t}|K_{j,t-1}=\upsilon_i) \end{aligned}$$

for $i=0,1$, where we have seen that $\omega_{j,k,i}=(1-\omega_{j,i,i})$. Thus, it turns out that the full conditionals for $\omega_{j,1,1}$ and $\omega_{j,0,0}$ are

$$\begin{aligned} (\omega_{j,1,1}|\boldsymbol{\Theta}_{\backslash\omega_{j,1,1}},\boldsymbol{y}) &\sim \mathcal{B}(a_\omega+\#\{t:\upsilon_1\to\upsilon_1\},b_\omega+\#\{t:\upsilon_1\to\upsilon_0\}), \\ (\omega_{j,0,0}|\boldsymbol{\Theta}_{\backslash\omega_{j,0,0}},\boldsymbol{y}) &\sim \mathcal{B}(a_\omega+\#\{t:\upsilon_0\to\upsilon_0\},b_\omega+\#\{t:\upsilon_0\to\upsilon_1\}), \end{aligned}$$

where $\#\{t:\upsilon_i\to\upsilon_k\}$ denotes the number of time points from $t=2$ to $t=T$ that $K_{j,t}$ switched from value υ_i to υ_k, $i,k=0,1$ with $\upsilon_1=1$ and $\upsilon_0=r$.

Finally, as already mentioned, the scaled states $\tilde{\boldsymbol{\beta}}_j=(\tilde{\beta}_{j,1},...,\tilde{\beta}_{j,T})$, $j=1,..,q$, can be jointly updated via FFBS algorithm. In order to gain computational efficiency, because of independency of $\boldsymbol{\beta}_1,...,\boldsymbol{\beta}_q$, we adopt the following strategy. From Equation (5.11), we have seen that for each $t=1,...,T$

$$y_t=\boldsymbol{F}_t'\tilde{\boldsymbol{\beta}}_t+u_t=\tilde{X}_{1,t}\tilde{\beta}_{1,t}+...+\tilde{X}_{q,t}\tilde{\beta}_{q,t}+u_t, \quad u_t\sim\mathcal{N}(0,1),$$

where $\tilde{X}_{j,t} = X_{j,t}\sqrt{\psi_{j,t}}$ and $\tilde{\beta}_{j,t} = \beta_{j,t}/\sqrt{\psi_{j,t}}$. Instead of working with matrix $\boldsymbol{F}_t'$ and with the system matrices $\boldsymbol{\Gamma}_t$ and $\boldsymbol{G}_t$, we can draw each $\boldsymbol{\beta}_j$ separately in a univariate FFBS as presented in Algorithm 7.

Algorithm 7: Univariate forward filtering backward sampling

```
For each iteration $m = 2, ..., M$:
1. Set $\tilde{\boldsymbol{\beta}}^{(m)} = \tilde{\boldsymbol{\beta}}^{(m-1)}$.
2. Then, for each $j = 1, .., q$:
a) Define $z_t = y_t - \sum_{i \neq j} \tilde{X}_{i,t}\tilde{\beta}_{i,t}^{(m)}$, such as $z_t = \tilde{X}_{j,t}\tilde{\beta}_{j,t} + u_t$ for $t = 1, ..., T$.
b) Draw $\tilde{\boldsymbol{\beta}}_j^{(m)}$ using the FFBS as in Algorithm 4 with the respective inputs for the
   Kalman filter and smoother.
```

Note that the inputs $\boldsymbol{\Gamma}_t, \boldsymbol{G}_t$ from (5.11), and the initial condition parameters for the states $\tilde{\beta}_{j,0} \sim \mathcal{N}(m_0, C_0)$, with $m_0 = E(\beta_{j,0})$ and $C_0 = 1$, are now scalars rather than matrices.

The same strategy is applied to update variables $\boldsymbol{K}_j = (K_{j,1}, ..., K_{j,T})$, $j = 1, .., q$, which can be jointly updated via Gerlach et al. (2000) algorithm showed in Algorithm 6 from Section 3.3.4. Because we have defined independent priors for $\boldsymbol{K}_1, ..., \boldsymbol{K}_q$ as in Definition 5.3, we can also define an univariate version of this algorithm for computational reasons. This strategy is summarized by the Algorithm 8 as follows.

Algorithm 8: Univariate version of the algorithm of Gerlach et al. (2000)

```
For each iteration $m = 2, ..., M$:
1. Set $\boldsymbol{K}^{(m)} = \boldsymbol{K}^{(m-1)}$.
2. Then, for each $j = 1, .., q$:
a) Define $z_t = y_t - \sum_{i \neq j} \tilde{X}_{i,t}\tilde{\beta}_{i,t}^{(m)}$, such as $z_t = \tilde{X}_{j,t}\tilde{\beta}_{j,t} + u_t$ for $t = 1, ..., T$.
b) Draw $\boldsymbol{K}_j^{(m)}$ using Algorithm 6 with the respective inputs.
```

5.2 Simulated examples

In this section we present two simulated examples where some coefficients are relevant in some periods of time and negligible in others. The first example is due to Kalli and Griffin (2014) and the second example is an application of the modified Cholesky decomposition where we simulate time-varying coefficients that compose the Cholesky factor $\boldsymbol{T}_t = (\boldsymbol{I} - \boldsymbol{\Phi}_t)$ and then apply the spike-and-slab priors on each successive regression.

5.2.1 First simulation example

We generated the data using Equation (5.1) with $q = 5$ predictors, $T = 200$ and constant observational variance $\sigma_t^2 = \sigma^2 = 1$, where $\boldsymbol{X}_t \sim N(\boldsymbol{0}, \boldsymbol{I})$ and $X_{j,1}, ..., X_{j,T}$ are independent. We simulate the five regression coefficients as follows.

1. The first coefficient $\beta_{1,t}$ follows a stationary AR(1) process with AR parameter 0.97 and a Normal stationary distribution with mean 2 and variance 0.25, i.e., for $t = 2, ..., T$

$$\beta_{1,t} = \mu_1 + 0.97\beta_{1,t-1} + \epsilon_{1,t}, \quad \epsilon_{1,t} \sim \mathcal{N}(0, \sigma_\epsilon^2),$$

where $\mathbb{E}(\beta_{1,t}) = \mathbb{E}(\beta_{1,t-1}) = \frac{\mu}{1-\phi_1} = 2$ and $\text{Var}(\beta_{1,t}) = \text{Var}(\beta_{1,t-1}) = \frac{\sigma_\epsilon^2}{1-\phi_1^2} = 0.25$. Thus, $\mu = 2(1-\phi_1) = 2(1-0.97) = 0.06$ and $\sigma_\epsilon^2 = 0.25(1-0.97^2) = 0.014775$. The initial value was drawn from its stationarity distribution $\beta_{1,1} \sim \mathcal{N}(2, 0.25)$.

2. The second coefficient $\beta_{2,t}$ also follows an AR(1) process with autocorrelation parameter 0.97 and a Normal marginal distribution with mean 0 and variance 0.25, but only until the half of the sample, that is:

$$\beta_{2,t} = \begin{cases} 0.97\beta_{2,t-1} + \epsilon_{2,t}, \quad \epsilon_{2,t} \sim \mathcal{N}(0, 0.014775), & t \leq 100 \\ 0, & t > 100, \end{cases}$$

with the initial value drawn as $\beta_{2,1} \sim \mathcal{N}(2, 0.25)$.

3. The third coefficient is always zero, except from two short periods when it equals -2:

$$\beta_{3,t} = \begin{cases} 0, & t \leq 20; 51 \leq t \leq 120; 151 \leq t \leq 200 \\ -2, & 21 \leq t \leq 50; 121 \leq t \leq 150. \end{cases}$$

4. The fourth coefficient is $\beta_{4,t} = 0, \forall t$.

5. The fifth coefficient $\beta_{5,t} = 0, \forall t$.

We sample from the posterior distribution using the three mentioned priors for $\beta_{j,t}$: dynamic NMIG, dynamic NG and dynamic Laplace with the the following hyperparameters settings: $\upsilon_0 = r = 0.005, \upsilon_1 = 1, a_\tau = 0.5$ (for the NG prior), $\nu = 15, c_0 = 51, C_0 = 5, a_\sigma = 0.0001, b_\sigma = 0.0001$ (improper prior) and $\alpha = 1000$ (tuning parameter for Metropolis). The MCMC algorithm was run for 20,000 iterations with half discarded as a burn-in.

The prior for autoregressive parameter is $\phi_j \sim \mathcal{B}(77.6, 2.4)$ for $j = 1, ..., 5$, so that it has mean 0.97. The same choice was made for the transition probabilities $\omega_{j,0,0}$ and $\omega_{j,1,1}$ so that both $\beta_{j,t}|\beta_{j,t-1}$ and $\psi_{j,t}|\psi_{j,t-1}$ evolves smoothly.

The RMSE results are shown in Table 5.1 and the fit of the models using the posterior median are shown in Figure 5.2.

Prior	RMSE (mean)	RMSE (median)
NMIG	0.3376	0.3268
NG	0.3678	0.3529
Laplace	0.3184	0.3144

Table 5.1: *RMSE for the dynamic spike-and-slab priors using the mean and the median of the sampled coefficients - simulated example 1*

We note that the dynamic Laplace prior was slightly superior in terms of RMSE than the other two priors. The dynamic NMIG prior also performs very similar, although we can see that both the dynamic Laplace and the dynamic NG have some issues: they are much more volatile than the NMIG prior. In fact, this is why we had to set a tight prior for Q_j. From (5.4), we see that

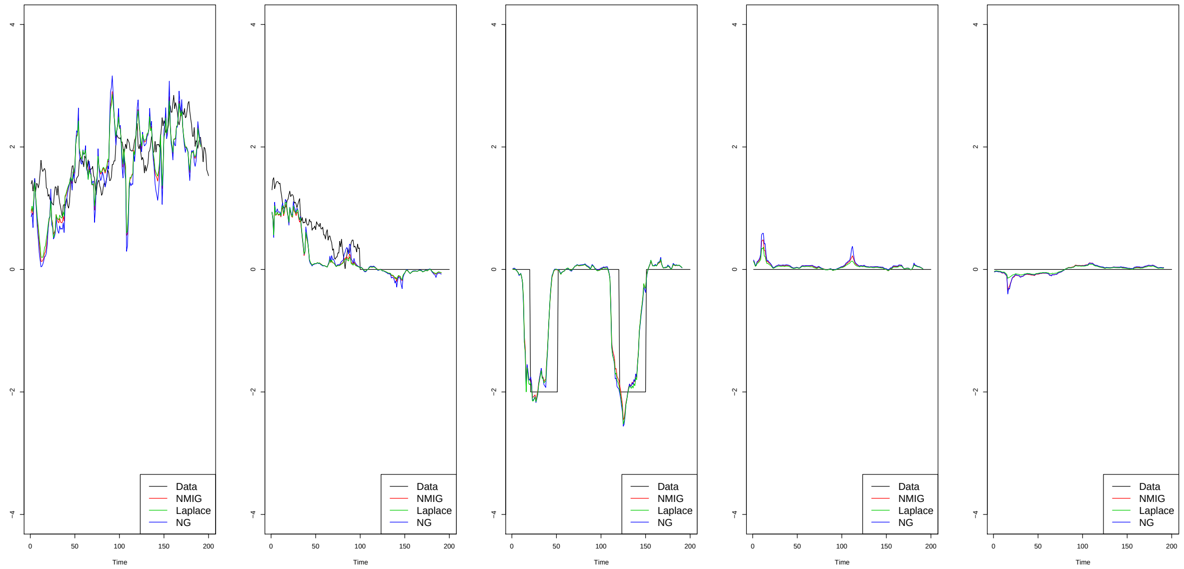


Figure 5.2: *Fit of the models using posterior medians*

$$\mathrm{Var}(Q_j) = \frac{(C_0/s^*(w))^2}{(c_0-1)^2(c_0-2)},$$

so that setting $c_0 = 51$ diminishes the variance. Another important thing to note is that we should give an appropriate value for C_0, depending on the data. We know that the prior variance of the slab at each time point t is

$$\mathrm{Var}_{slab}(\beta_{j,t}) = cQ_j,$$

where c is the constant that depends on the mixing density. For this data, the prior variance is equal 0.105 if $\omega_{j,1,i} = 0.95$ and is equal 0.957 if $\omega_{j,1,i} = 0.1$. This means that if there is no switch of regime (that is, probably $K_{j,t-1} = 1$, then $\omega_{j,1,i}$ is high), then the variance of the slab decreases.

Because of this issue of the volatility of the NG and Laplace posterior medians, we ran a test using a fixed value for $s^*(w) = c[0.5r + 0.5]$ and the following settings: $\upsilon_0 = r = 0.000025, \upsilon_1 = 1, a_\tau = 0.5$ (for the NG prior), $\nu = 5, c_0 = 4, C_0 = 0.5, a_\sigma = 0.0001, b_\sigma = 0.0001$ (improper prior) and $\alpha = 1000$ (tuning parameter for Metropolis). The MCMC algorithm was run for 30,000 iterations with half discarded as a burn-in.

The RMSE results are shown in Table 5.2 and the fit of the models using the posterior median are shown in Figure 5.2.

One can see that in general the RMSE is lower than the previous setting. The Laplace prior still has the lowest RMSE. Even though the RMSE is lower, we also noticed that for the third coefficient, where there is a sudden change in its value, the flexible structure that allows $s^*(w)$ to depend on the value of $\omega_{j,1,i}$ seems to adapt better.

Prior	RMSE (mean)	RMSE (median)
NMIG	0.3057	0.3178
NG	0.2996	0.3091
Laplace	0.2986	0.3022

Table 5.2: *RMSE for the dynamic spike-and-slab priors using the mean and the median of the sampled coefficients - simulated example 1 with fixed $s^*(w)$*

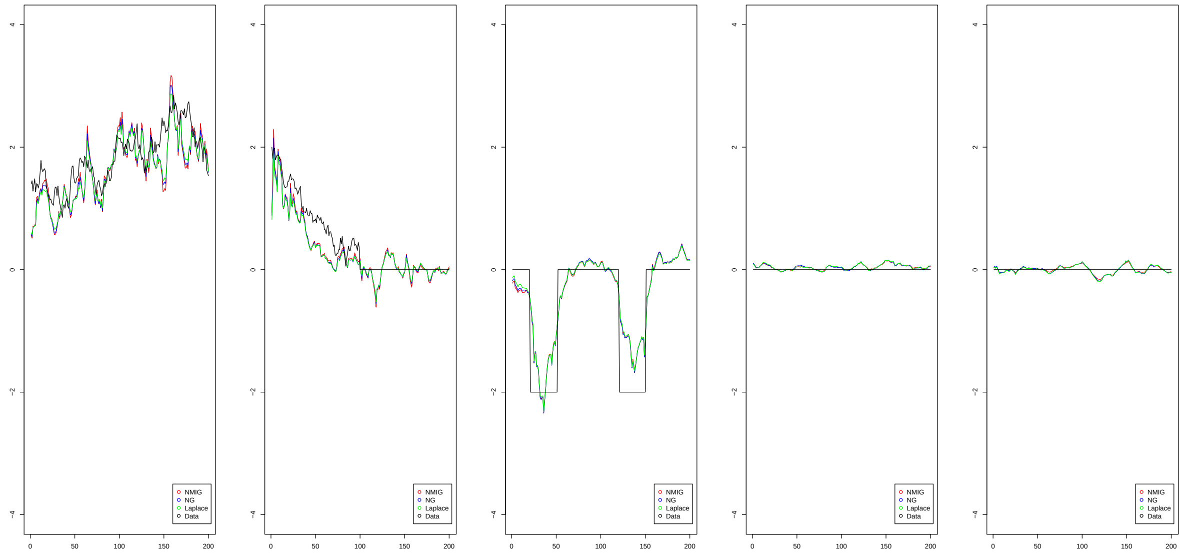


Figure 5.3: *Fit of the models using posterior medians - fixed $s^*(w)$*

5.2.2 Second simulation example

In this second example we simulate time-varying coefficients that compose the Cholesky factor $\boldsymbol{T}_t$ and then apply the spike-and-slab priors on each successive regression. The simulation is done as follows. We define that the number of time points $T = 240$ and the number of ordered variables that compose the vector $\boldsymbol{y}$ is $q = 10$.

In the Cholesky decomposition each variable is regressed on its predecessors in a dynamic regression problem. We know by Equation (4.8) from Chapter 4 that

$$y_{i,t} = \sum_{j=1}^{i-1} \beta_{i,j,t} y_{j,t} + \varepsilon_{i,t}$$

for $i = 2, .., q$, with $y_{1,t} = \varepsilon_{1,t}$.

The Cholesky factor is then

$$\boldsymbol{T}_t = (\boldsymbol{I} - \boldsymbol{B}_t),$$

where $\boldsymbol{B}_t$ is the lower triangular matrix of coefficients for each time t with zeros in the diagonal, that is, the matrix with entries $\beta_{2,1,t}, \beta_{3,1,t}, \beta_{3,2,t}, ..., \beta_{q,1,t}, ..., \beta_{q,q-1,t}$. Thus, we have $q(q-1)T$ parameters to be estimated.

We define four possible processes for the time-varying coefficients and then sample from these possibilities using predetermined probabilities.

1. A stationary AR(1) process with autoregressive coefficient $\phi = 0.98$ and with fixed variance $\sigma^2 = (1 - \phi)0.15$, without an intercept term, that is

$$\beta_{i,j,t} = \phi\beta_{i,j,t-1} + \nu_{i,j,t},$$

with $\nu_{i,j,t} \sim \mathcal{N}(0, \sigma^2)$.

2. A stationary AR(1) process with autoregressive coefficient $\phi = 0.98$ and with fixed variance $\sigma^2 = (1 - \phi)0.15$ until the half of the time points. Then, the coefficient is set to zero.

3. A fixed interval process similar to the third coefficient from the first simulated example as follows

$$\beta_{i,j,t} = \begin{cases} 0, & t \leq T/8; 3T/8 < t \leq 5T/8; t > 7T/8 \\ -0.5, & T/8 \leq t < 3T/8; 5T/8 < t \leq 7T/8. \end{cases}$$

4. A constant coefficient equal to zero.

In this manner, we want to give a structure to the Cholesky factor, but now allowing for time-varying coefficients. Each coefficient $\beta_{i,j,t}$ follows one of first three processes: (1) AR(1), (2) AR(1) with zeros, (3) fixed intervals, or (4) zeros, which are sampled using equal probabilities. Then we build the 10 time series $\boldsymbol{y}_1, ..., \boldsymbol{y}_{10}$ as

$$\begin{aligned} y_{1,t} &= \varepsilon_{1,t} \\ y_{2,t} &= \beta_{2,1,t} y_{1,t} + \varepsilon_{2,t} \\ &\dots \\ y_{10,t} &= \sum_{j=1}^{9} \beta_{10,j,t} y_{j,t} + \varepsilon_{10,t}, \end{aligned}$$

for $t = 1, .., 240$ and where $\varepsilon_{i,t} \sim \mathcal{N}(0, 0.0625)$, $\forall j = 1, .., 10$.

The aim of these simulation is to see if the errors accumulate because now we are dealing with 10 variables and 9 individual equations indexed by i, each one with $j = 1, ..., (i - 1)$ regressors. The results for the RMSE are shown in Table 5.3. The fit of the coefficients for the last equations ($i = 10$) of each of the three priors (NMIG, NG, Laplace) is presented in Figures 5.4, 5.5, 5.6, respectively. The MCMC scheme uses 10,000 simulations with $5,000$ discarded as burn-in. The hyperparameters were set as follows: $\upsilon_0 = r = 0.005, \upsilon_1 = 1, a_\tau = 0.5$ (for the NG prior), $\nu = 25, c_0 = 50, C_0 = 1.5, a_\sigma = 5, b_\sigma = 1.5$ and $\alpha = 1000$ (tuning parameter for Metropolis).

We have noticed that for this data and settings, the RMSE of three models are very similar. Indeed, the 95 % confidence intervals are also very similar for all priors, and we notice that the fit is quite satisfactory.

Prior	RMSE (mean)	RMSE (median)
NMIG	0.2472	0.2863
NG	0.2398	0.2820
Laplace	0.2401	0.2842

Table 5.3: *RMSE for the dynamic spike-and-slab priors using the mean and the median of the sampled coefficients - simulated example 2*

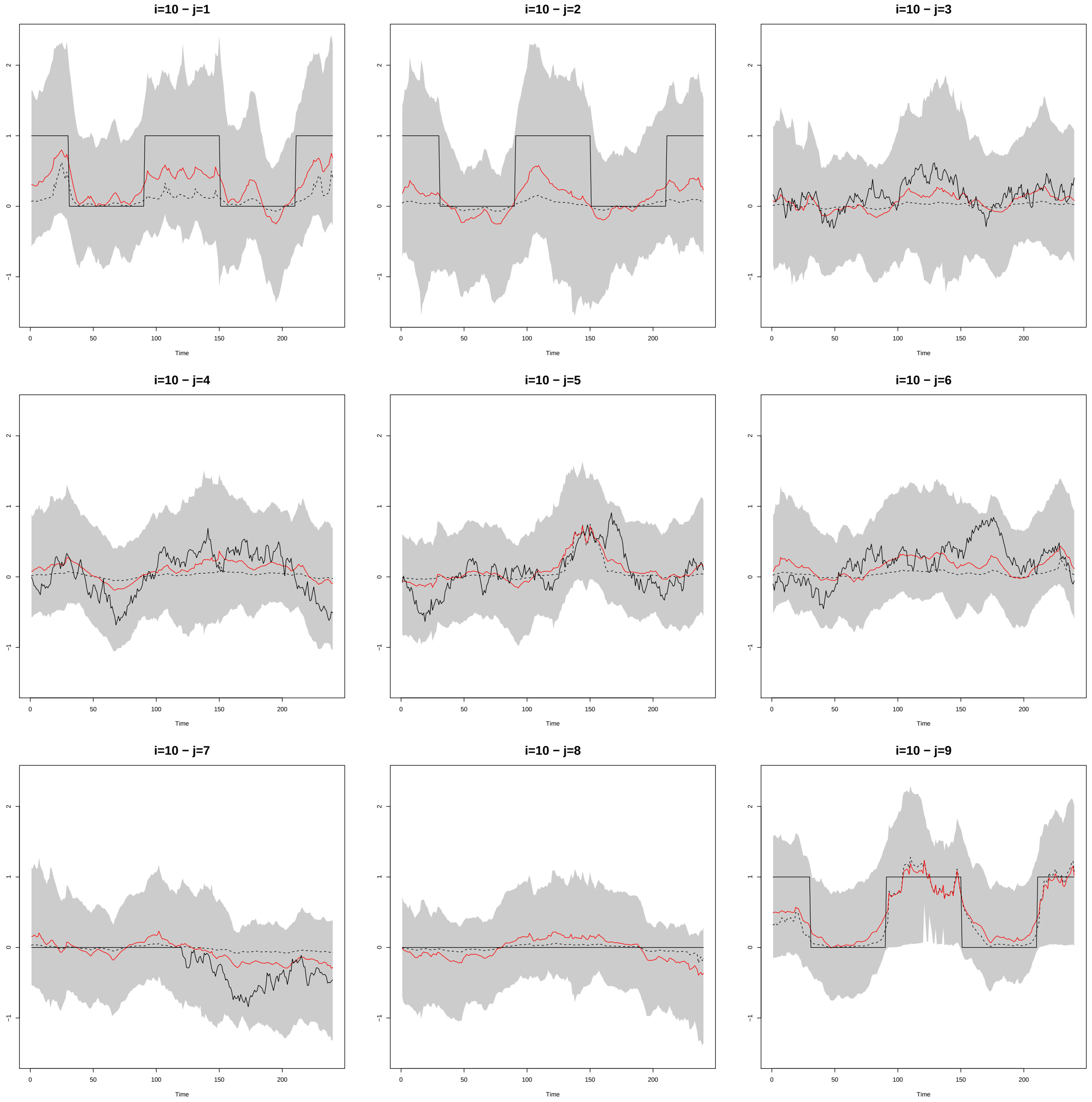


Figure 5.4: *95% confidence intervals of the dynamic NMIG prior. Median=dashed line; Real data=black; Posterior mean=red.*

5.3 Empirical example

The empirical application is due to Kalli and Griffin (2014) and uses inflation data obtained from Professor Griffin's research page[1]. We use the inflation data collected by them with the

[1] Available in https://www.kent.ac.uk/smsas/personal/jeg28/index.htm

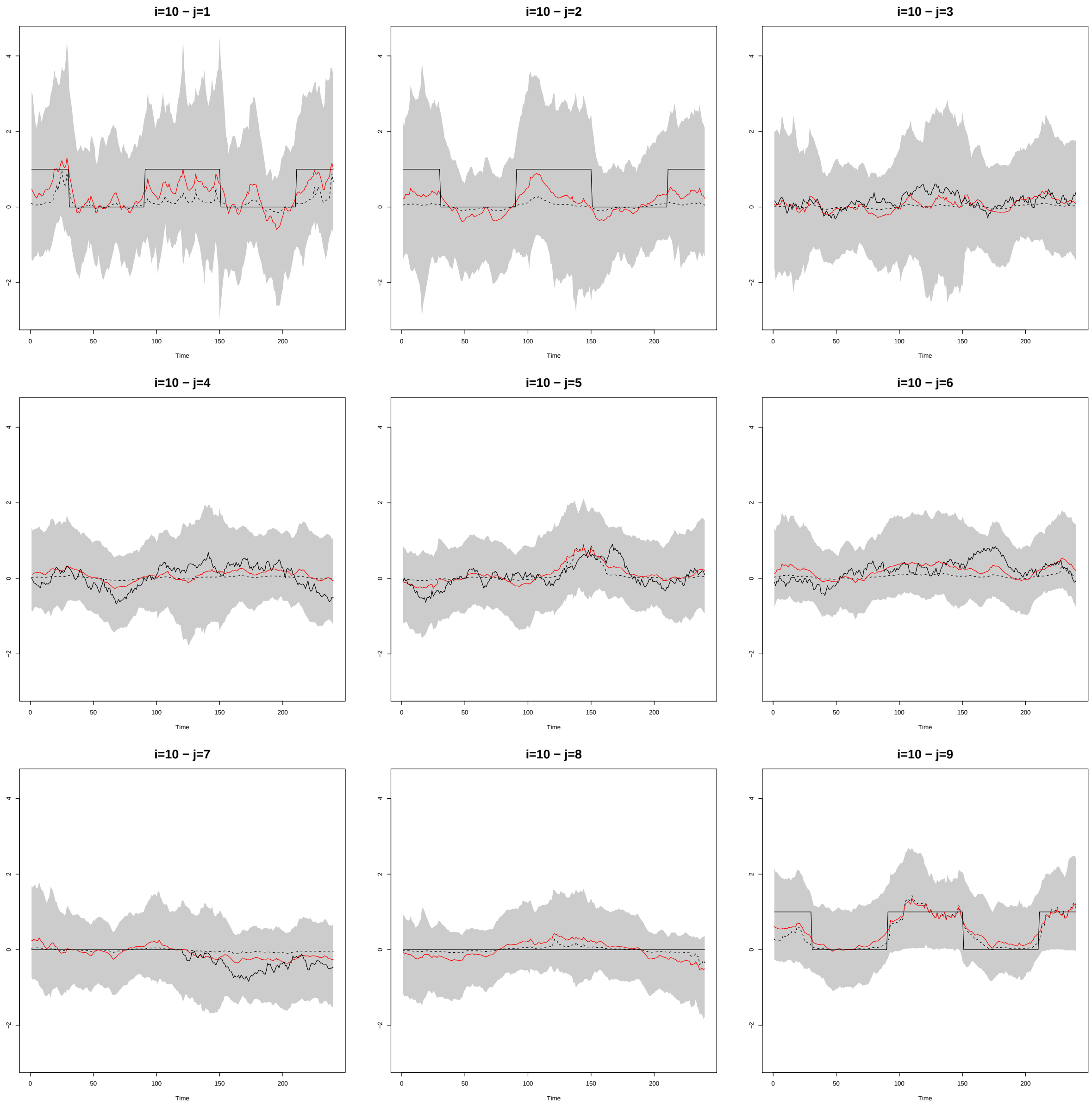


Figure 5.5: *95% confidence intervals of the dynamic NG prior.*
Median=dashed line; Real data=black; Posterior mean=red.

independent variable as the US quarterly inflation measure based on the Gross Domestic Product (GDP). The data was obtained from FRED database, Federal Reserve Bank of St.Louis, University of Michigan Consumer Survey database, Federal Reserve Bank of Philadelphia, and Institute of Supply Management. The data set includes 31 predictors, from activity and term structure variables to survey forecasts and previous lags. A full description of the 31 explanatory variables can be found in Appendix. The sample period is from the second quarter of 1965 to first quarter of 2011 with $T = 182$ observations.

Inflation forecasting is a frequent topic within the shrinkage in time varying parameter models literature and was also the main subject of Belmonte et al. (2014). The size of the set of potential variables to forecast inflation is huge and, as noted by Kalli and Griffin (2014), this is usually split into four subsets: past inflation forecasts, where the explanatory variables are previous lags of inflation; Phillips curve forecasts, which involve activity variables, such as economic growth

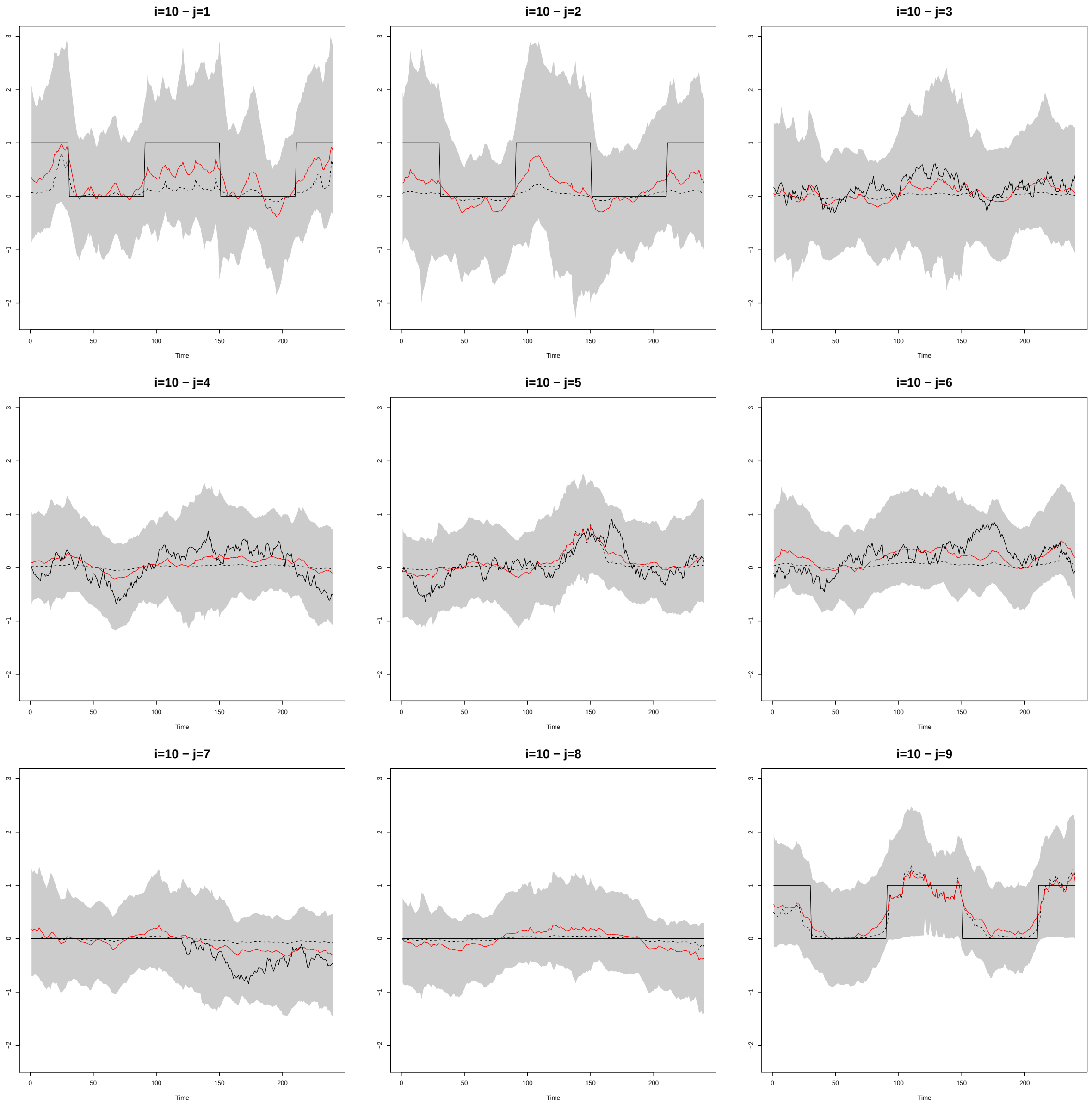


Figure 5.6: *95% confidence intervals of the dynamic Laplace prior. Median=dashed line; Real data=black; Posterior mean=red.*

rate or output gap, unemployment rate, and lagged inflation; forecasts based on variables which are themselves forecasts of asset prices (combination indices), term structures of nominal debt, and consumer surveys; and forecasts based on other exogenous variables such as government investment, the number of new private houses.

We applied the three dynamic variable selection priors to the GDP deflator data with the following hyperparameter settings: $v_0 = r = 0.05, v_1 = 1, a_\tau = 0.5$ (for the NG prior), $\nu = 50, c_0 = 50, C_0 = .05, a_\sigma = 31, b_\sigma = 30\hat{\sigma}^2 = 4.22$, with $\hat{\sigma}^2 = 0.14$ being the sum of square residuals of the OLS estimate divided by $(T-1)$ and $\alpha = 1000$ (tuning parameter for Metropolis). The previous Beta priors, that is, $\phi_j \sim \mathcal{B}(77.6, 2.4)$ and for the transition probabilities $\omega_{j,0,0} \sim \mathcal{B}(77.6, 2.4)$ and $\omega_{j,1,1} \sim \mathcal{B}(77.6, 2.4)$ were maintained. We ran a total of 20,000 iterations of the MCMC scheme and we discarded 10,000 as a burn-in.

The results (the mean of the coefficients $\beta_{j,t}$ and the relevances $\psi_{j,t}$ were compared to results from the NGAR process defined Kalli and Griffin (2014). We used the MATLAB code provided by Professor Griffin in his website for the GDP inflation data after standardizing both the response and the predictors in the same way as done by the authors. In Figures 5.7, 5.8 and 5.9 we compare the coefficients and in Figures 5.10, 5.11 and 5.12 we compare the relevances for the 16 predictors highlighted by the authors in their paper.

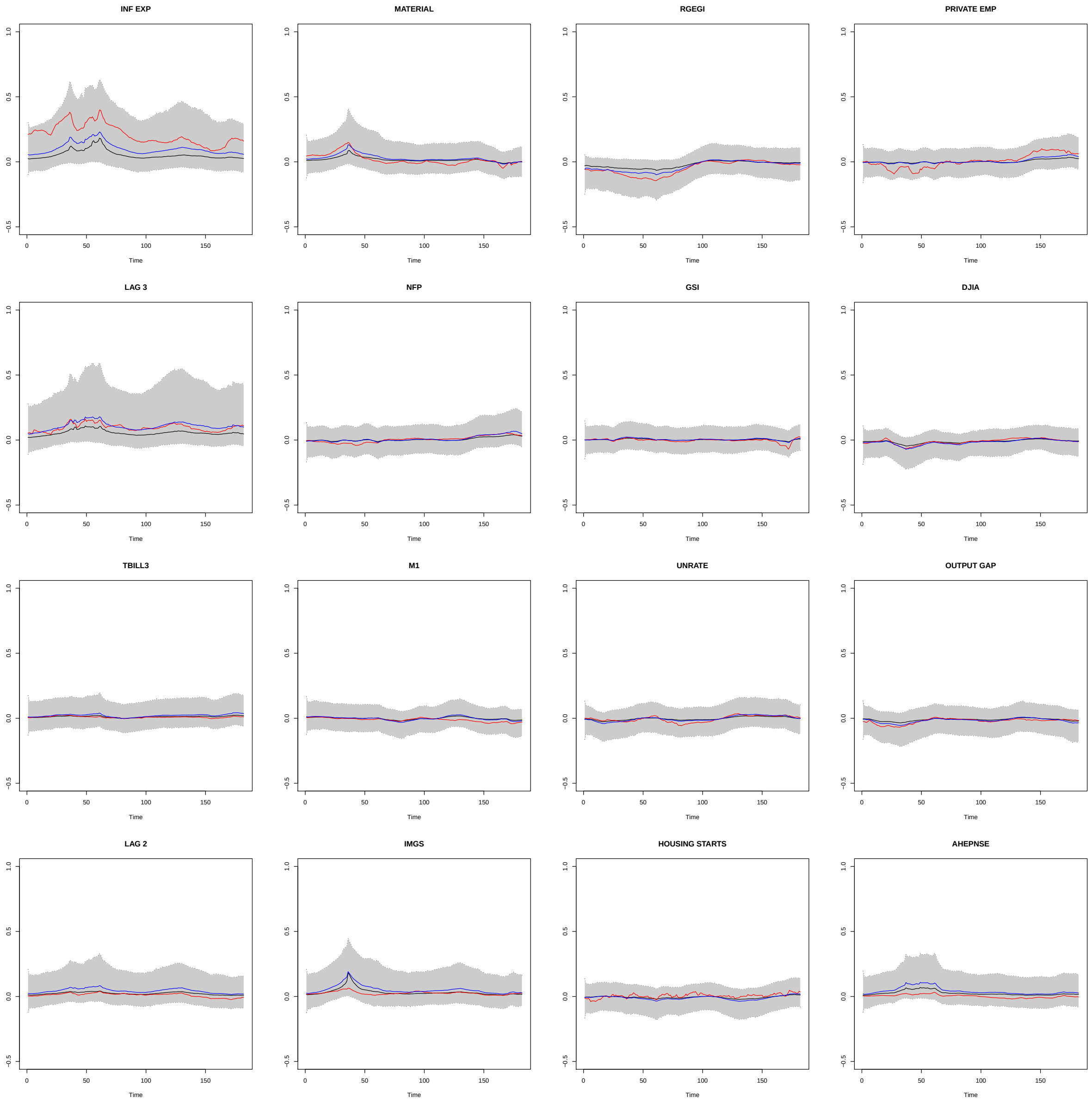


Figure 5.7: *Comparison between the mean of the NGAR model and the dynamic NMIG prior coefficients. Mean NGAR=red line; Mean NMIG=blue line; Median NMIG=black line; 95% confidence intervals=grey area.*

In general, the coefficients are quite similar and the means of the sampled coefficients from the NGAR process is inside our prior's confidence interval except from the expected inflation (INF EXP) which is slightly higher for the NGAR process. We can see that INF EXP is clearly an important predictor of the GDP deflator. It is more important in the mid 1970's and mid 1980's. Its coefficient is positive from the start of our sample period up to the start of the 2000s, when it starts to approach zero. The lower band of its 95% CI suggests that it may also have a

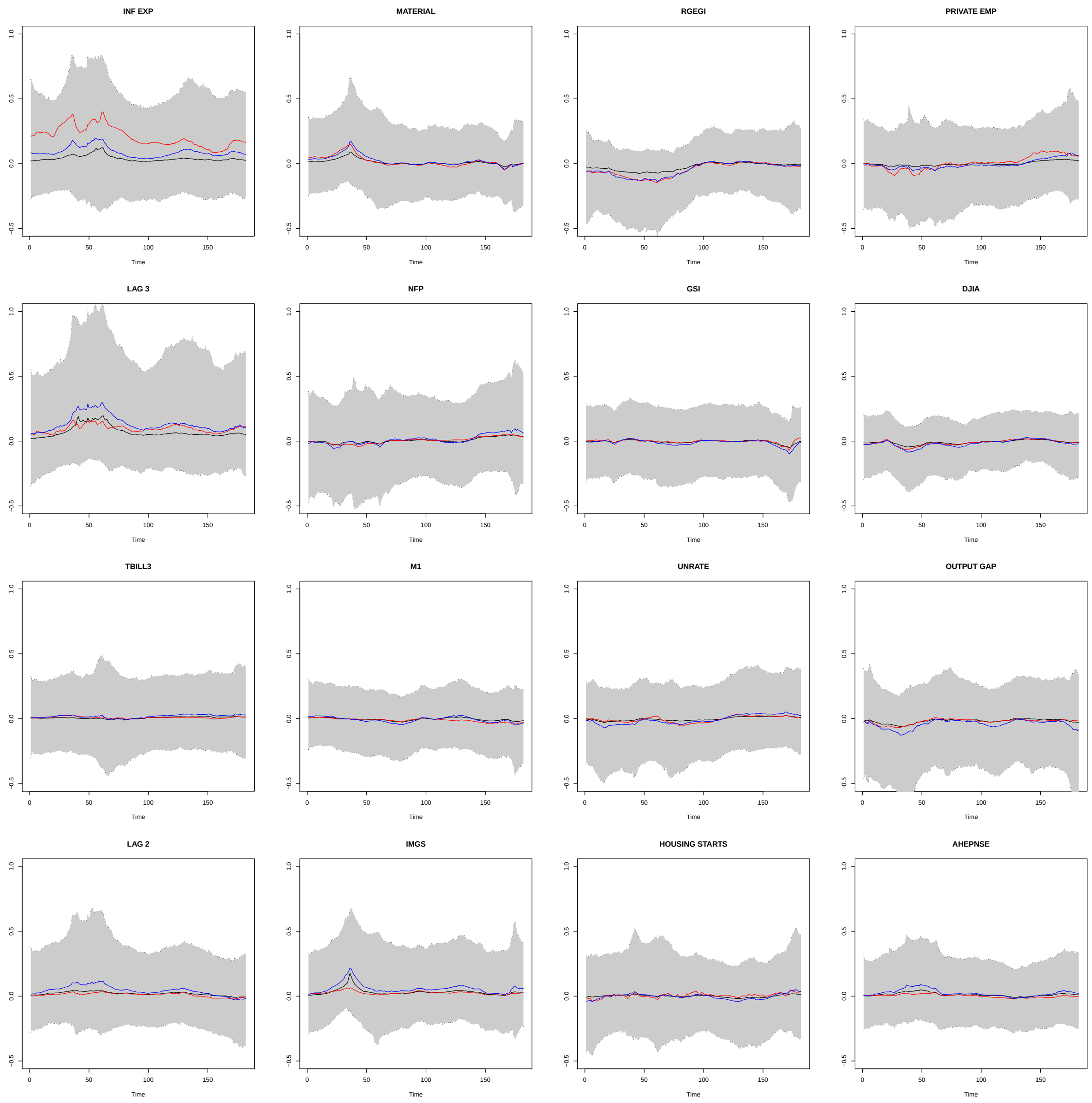


Figure 5.8: *Comparison between the mean of the NGAR model and the dynamic NG prior coefficients. Mean NGAR=red line; Mean NG=blue line; Median NG=black line; 95% confidence intervals=grey area.*

negative effect on the GDP deflator. The coefficient of RGEGI growth is negative in the 1980's, however its effect on the GDP deflator is clearly more obvious.

Lastly, it is important to note that Laplace and NG prior result in wider confidence interval for the coefficients than the NMIG prior relative to the NGAR process.

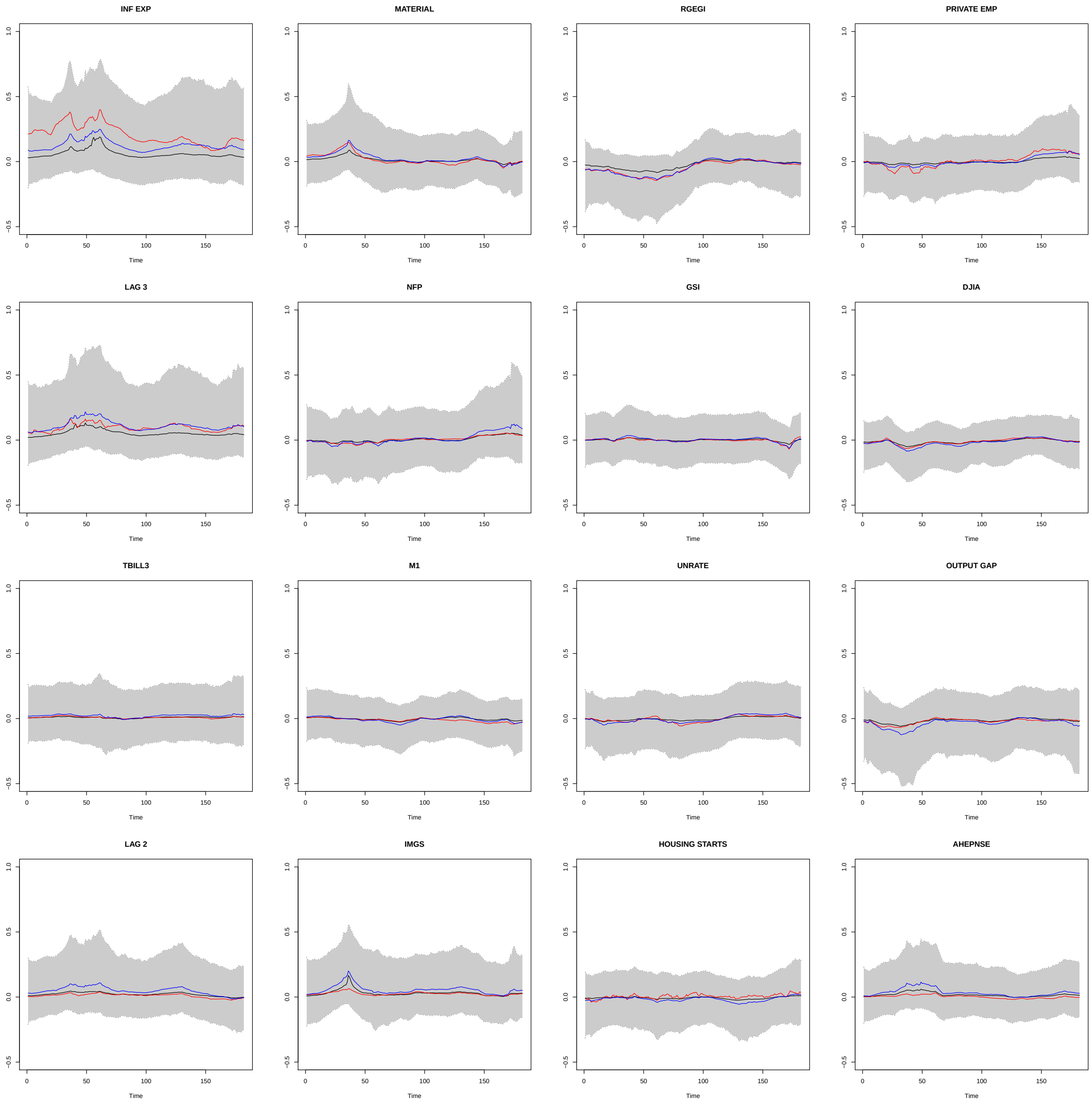


Figure 5.9: *Comparison between the mean of the NGAR model and the dynamic Laplace prior coefficients. Mean NGAR=red line; Mean Laplace=blue line; Median Laplace=black line; 95% confidence intervals=grey area.*

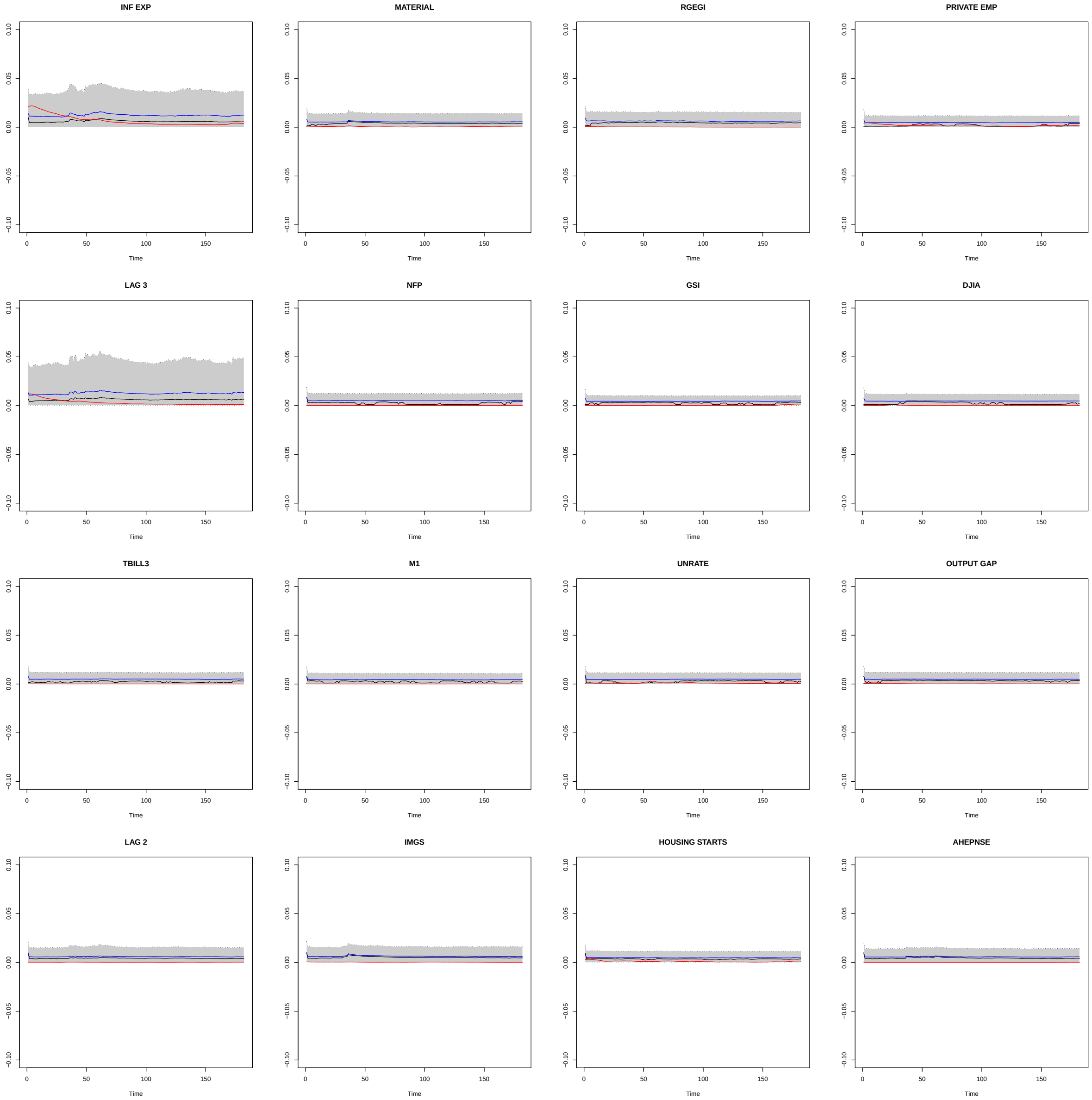


Figure 5.10: *Comparison between the mean of the NGAR model and the dynamic NMIG prior relevances. Mean NGAR=red line; Mean NMIG=blue line; Median NMIG=black line; 95% confidence intervals=grey area.*

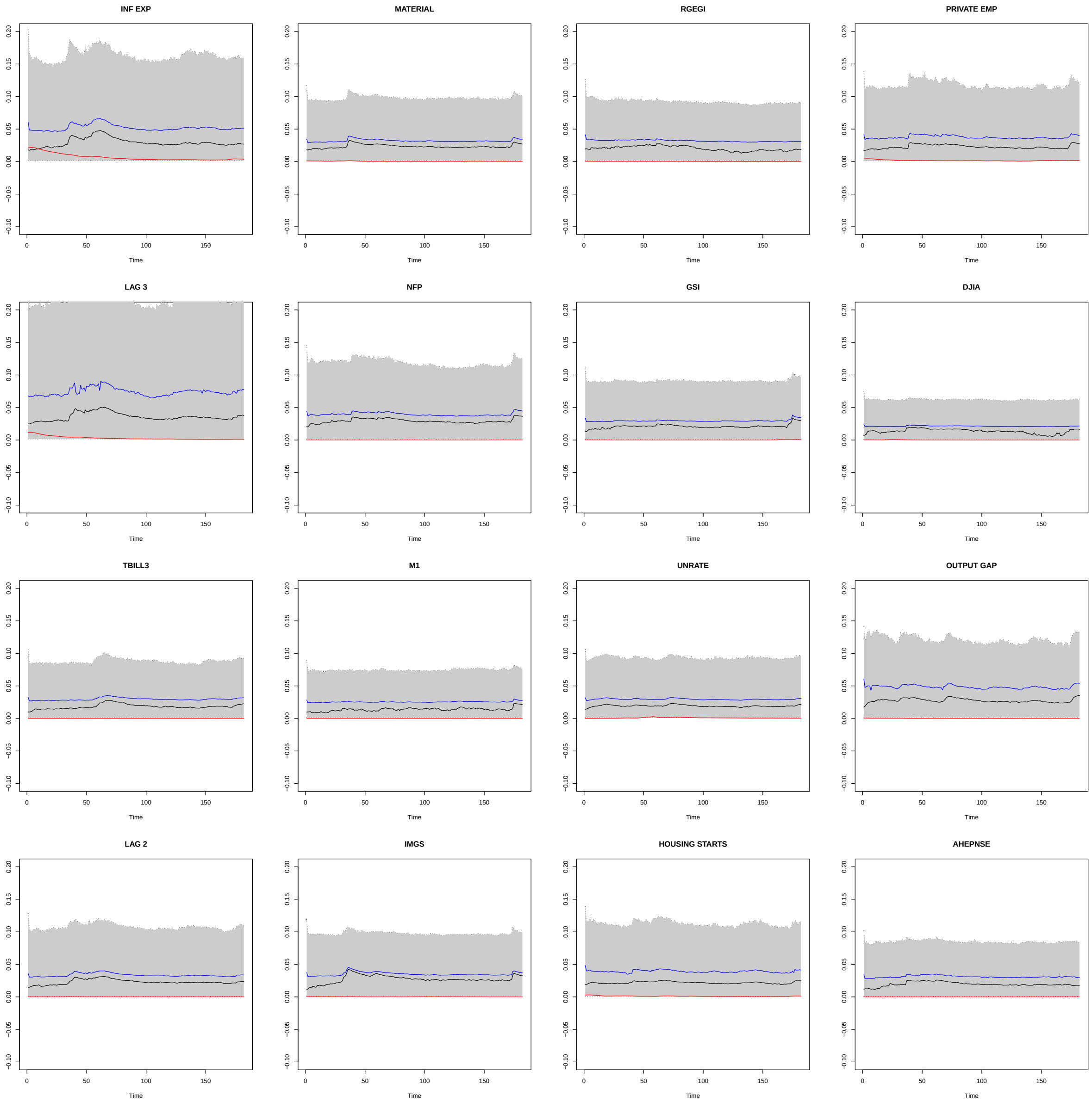


Figure 5.11: *Comparison between the mean of the NGAR model and the dynamic NG prior relevances. Mean NGAR=red line; Mean NG=blue line; Median NG=black line; 95% confidence intervals=grey area.*

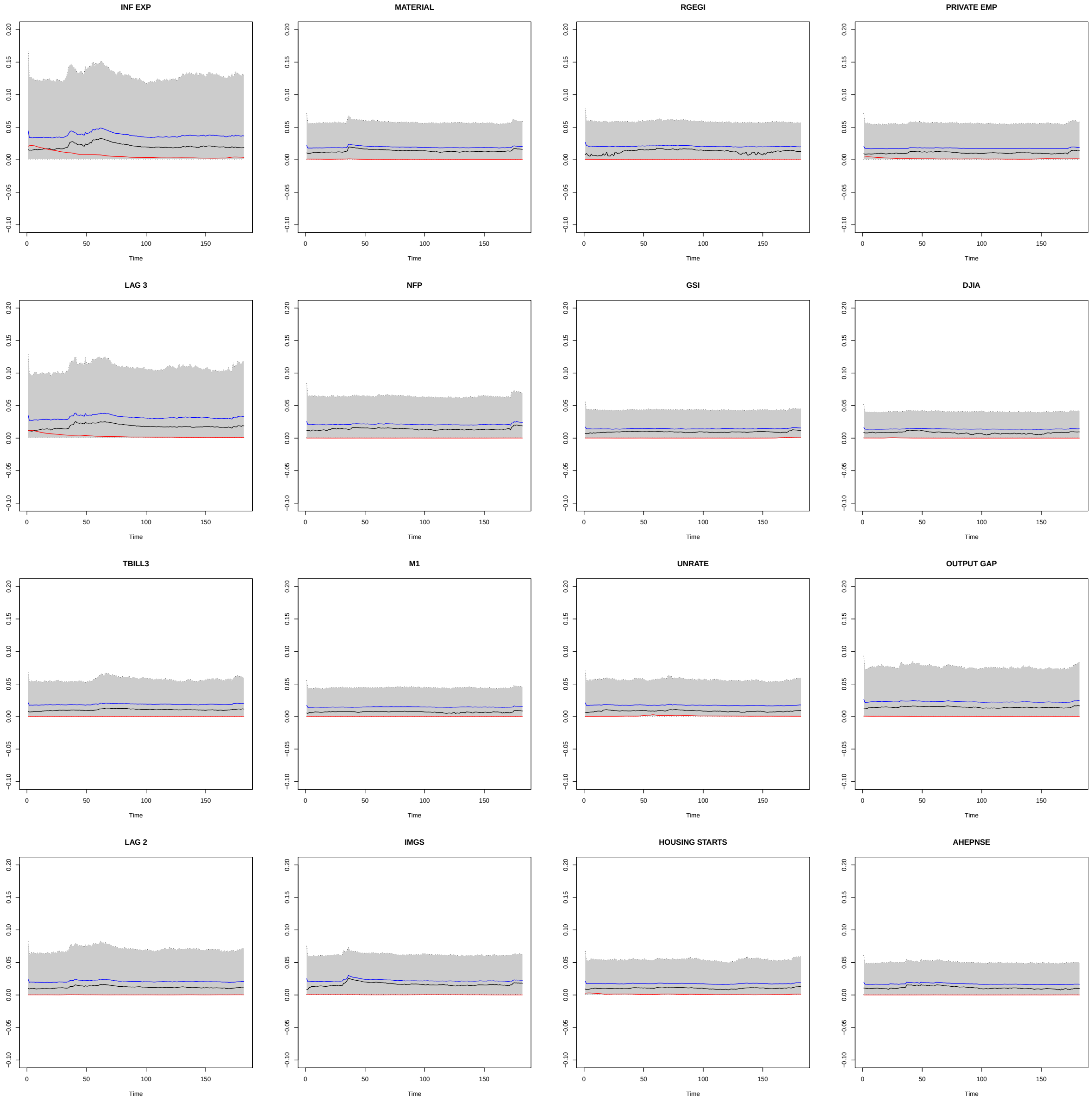


Figure 5.12: *Comparison between the mean of the NGAR model and the dynamic Laplace prior relevances. Mean NGAR=red line; Mean Laplace=blue line; Median Laplace=black line; 95% confidence intervals=grey area.*

Chapter 6

Conclusion

In this work, we have discussed several variable selection and regularization methods within the linear regression framework. The modified Cholesky decomposition was presented as a simple and with good interpretation way to decompose covariance matrices in linear regression problems. In particular, we showed some existing methods to regularize the Cholesky factor, that is, the lower triangular matrix composed by the coefficients of the regressing each variable on its predecessors (possibly considering a natural order for the variables). Then we applied the Normal-Gamma prior on the coefficients and compare with these methods using finance data of returns from the S&P 100 index.

We also developed a regularization model for Gaussian dynamic regression problems in which the sparsity pattern varies over time. Our method is somewhat similar to references like Belmonte et al. (2014) and Bitto and Frühwirth-Schnatter (2016) as we have used the scaled coefficients $\tilde{\beta}_{j,t} = \beta_{j,t}/\sqrt{\psi_{j,t}}$ in the Gaussian DLM state equation, thus arising to what they have called non-centered parametrization. Then, the shrinkage of the time-varying coefficients $\beta_{j,t}$ was done by assigning priors to the scale parameters $\psi_{j,t}$. The main difference between these two references and our approach is that we adopt a time-varying scaling parameter $\psi_{j,t}$ instead of using fixed over time standard deviations ω_j as the mentioned references did.

In both approaches the standard deviation ω_j plays the role of relevance of the jth predictor: small values of ω_j leads to greater shrinkage of the coefficient $\beta_{j,t}$ for all times t. In this sense, both approaches assigns horizontal sparsity, as the shrinkage effect of the prior for ω_j is equal over all times t. While Belmonte et al. (2014) used the Laplace prior for shrinking the standard deviations, Bitto and Frühwirth-Schnatter (2016) used the Normal-Gamma prior, which is more general since the Laplace prior is a special case from the Normal-Gamma prior where the shrinking parameter equals one.

Thus, our method is more general as it allows for vertical sparsity, i.e., by defining time-varying scale parameters we can have some periods of time that a predictor becomes irrelevant and in others its coefficients is non-zero. In that way, our model is closer to Kalli and Griffin (2014) NGAR process from Kalli and Griffin (2014). Nevertheless, we assume a Markov switching structure for $\psi_{j,t}$ instead of the GAR process. In fact, $\psi_{j,t}$ is a time-varying mixture process where the mixture weights $\omega_{j,1,i}$ change over time. This is the reason why we have called our prior *dynamic spike-and-slab prior*.

Through the introduction of latent Markov switching variables $K_{j,t}$, it was possible to vary the sparsity structure, so that at each instant of time the scale parameters $\psi_{j,t} = K_{j,t}\tau_j^2$

could assume a different mixture distribution. We allow not only the weights $\omega_{j,1,i}$ of the finite mixture to change, depending on the previous value of the latent variable, that is, $\omega_{j,1,i} = p(K_{j,t} = \upsilon_1 | K_{j,t-1} = \upsilon_i)$, but also the parameter Q_j of the distribution of τ_j^2 to change, depending on the value of $\omega_{j,1,i}$. The parameter Q_j is given an Inverted Gamma prior and also depends on a constant c which varies according to the mixing distribution chosen. Following Frühwirth-Schnatter and Wagner (2011), we test three priors for the coefficients $\beta_{j,t}$: the NMIG prior, the Normal-Gamma (NG) and the Laplace mixture prior or equivalently, three mixing distributions for τ_j^2: the Inverse-Gamma, the Gamma and the Exponential distribution, respectively.

Posterior inference is done by adopting a MCMC scheme: an hybrid Gibbs Sampler, where the scaled states $\tilde{\beta}_{j,t}$ are drawn using the FFBS algorithm and the latent variables $K_{j,t}$ are sampled using Gerlach et al. (2000) algorithm for Markov processes of order 1.

To exemplify the proposed model, we applied the three priors on simulated data where some coefficients become irrelevant at some times. The first simulation example is due to Kalli and Griffin (2014) and the second simulation is an exercise of using the dynamic spike-and-slab prior on the coefficients of the Cholesky regressions. The third application is a real data example with US inflation data.

From the results presented, we can conclude that the model works well for the simulated Cholesky factor and that the Laplace prior is slightly better in terms of RMSE. We notice the importance of setting the right hyperparameter values, specially for those related to parameter Q_j as we do not want the prior for Q_j to have a large variance. In practice, the value should be proportional to the magnitude of the coefficients even if the covariates are standardized.

Considering these, further research include the following topics:

- Test a Gamma prior for the parameter Q_j instead of the Inverted-Gamma or a Normal prior for $\sqrt{Q_j}$, following the work of Frühwirth-Schnatter and Wagner (2010) which criticizes the use of the Inverse-Gamma because the posterior values are strongly influenced by the hyperparameters.

- Construct other mixture priors such as a dynamic mixture of Student's-t and Laplace densities for the coefficients.

- Allow for time-varying observational variances using stochastic volatility models.

- Study time-varying sparsity for factorial models.

- Compare predictive performance with other existing methods.

Appendix

Name	Description
GDP	Difference in logs of real gross domestic product
PCE	Difference in logs of real personal consumption expenditure
GPI	Difference in logs of real gross private investment
RGEGI	Difference in logs of real government consumption expenditure and gross investment
IMGS	Difference in logs of imports of goods and services
NFP	Difference in logs non-farm payroll
M2	Difference in logs M2 (commercial bank money)
ENERGY	Difference in logs of oil price index
FOOD	Difference in logs of food price index
MATERIALS	Difference in logs of producer price index (PPI) industrial commodities
OUTPUT GAP	Difference in logs of potential GDP level
GS10	Difference in logs of 10yr Treasury constant maturity rate
GS5	Difference in logs of 5yr Treasury constant maturity rate
GS3	Difference in logs 3yr Treasury constant maturity rate
GS1	Difference in logs 1yr Treasury constant maturity rate
PRIVATE EMPLOYMENT	Log difference in total private employment
PMI MANU	Log difference in PMI-manufacturing index
AHEPNSE	Log difference in average hourly earnings of private non management employees
DJIA	Log difference in Dow Jones Industrial Average Returns
M1	Log difference in M1 (narrow-commercial bank money)
ISM SDI	Institute for Supply Management (ISM) Supplier Deliveries Inventory
CONSUMER	University of Michigan consumer sentiment (level)
UNRATE	Log of the unemployment rate
TBILL3 3m	Treasury bill rate
TBILL SPREAD	Difference between GS10 and TBILL3
HOUSING STARTS	Private housing (in thousands of units)
INF EXP	University of Michigan inflation expectations (level)
LAG1, LAG2, LAG3, LAG4	The first, second, third and fourth lag

Table 6.1: *Inflation Data. Sources: FRED database, Federal Reserve Bank of St.Louis, University of Michigan Consumer Survey database, Federal Reserve Bank of Philadelphia, and Institute of Supply Management.*

Bibliography

Hirotugu Akaike. Markovian representation of stochastic processes and its application to the analysis of autoregressive moving average processes. *Annals of the Institute of Statistical Mathematics*, 26(1):363–387, 1974. 29

Miguel AG Belmonte, Gary Koop, and Dimitris Korobilis. Hierarchical shrinkage in time-varying parameter models. *Journal of Forecasting*, 33(1):80–94, 2014. 40, 44, 46, 47, 48, 56, 69, 84, 93

Angela Bitto and Sylvia Frühwirth-Schnatter. Achieving shrinkage in a time-varying parameter model framework. *arXiv preprint arXiv:1611.01310*, 2016. 17, 40, 46, 48, 69, 93

Chris K Carter and Robert Kohn. On gibbs sampling for state space models. *Biometrika*, 81 (3):541–553, 1994. 3, 39

Joshua CC Chan, Gary Koop, Roberto Leon-Gonzalez, and Rodney W Strachan. Time varying dimension models. *Journal of Business & Economic Statistics*, 30(3):358–367, 2012. 40, 48, 51, 56, 69

Sylvia Frühwirth-Schnatter. Data augmentation and dynamic linear models. *Journal of time series analysis*, 15(2):183–202, 1994. 3, 39

Sylvia Frühwirth-Schnatter and Helga Wagner. Stochastic model specification search for gaussian and partial non-gaussian state space models. *Journal of Econometrics*, 154(1):85–100, 2010. 15, 40, 43, 45, 46, 47, 94

Sylvia Frühwirth-Schnatter and Helga Wagner. Bayesian variable selection for random intercept modeling of gaussian and non-gaussian data. *Bayesian Statistics 9*, 9:165, 2011. 4, 7, 18, 26, 27, 94

Dani Gamerman and Hedibert F Lopes. *Markov chain Monte Carlo: stochastic simulation for Bayesian inference.* CRC Press, 2006. 38

Edward I George and Robert E McCulloch. Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993. 2, 4, 7, 18, 19, 20

Edward I George and Robert E McCulloch. Approaches for bayesian variable selection. *Statistica sinica*, pages 339–373, 1997. 20

Richard Gerlach, Chris Carter, and Robert Kohn. Efficient bayesian inference for dynamic mixture models. *Journal of the American Statistical Association*, 95(451):819–828, 2000. iii, v, xv, 3, 48, 49, 51, 52, 54, 72, 73, 78, 94

Christian Gourieroux and Joann Jasiak. Autoregressive gamma processes. *Journal of Forecasting*, 25(2):129–152, 2006. 54

Jim E Griffin, Philip J Brown, et al. Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, 5(1):171–188, 2010. iii, v, 3, 14, 16, 18, 46, 53

P. J. Harrison and C. F. Stevens. Bayesian forecasting (c/r: p228-247). *Journal of the Royal Statistical Society, Series B: Methodological*, 38:205–228, 1976. 29

AC Harvey. Forecasting, structural time series models and the kalman filter. 1989. 29

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, 2001. 1, 2

Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970. 1, 4, 7, 11

Jianhua Z Huang, Naiping Liu, Mohsen Pourahmadi, and Linxu Liu. Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, 93(1):85–98, 2006. 4, 59, 61, 62, 63, 64, 65, 67

Hemant Ishwaran and J Sunil Rao. Spike and slab variable selection: frequentist and bayesian strategies. *Annals of Statistics*, pages 730–773, 2005. iii, v, 3, 4, 7, 18, 22, 23, 24, 26, 70

Eric Jacquier, Nicholas G Polson, and Peter E Rossi. Bayesian analysis of stochastic volatility models. *Journal of Business & Economic Statistics*, 1994. 46

Maria Kalli and Jim E Griffin. Time-varying sparsity in dynamic regression models. *Journal of Econometrics*, 178(2):779–793, 2014. iii, v, 40, 53, 56, 69, 72, 78, 83, 84, 86, 93, 94

Rudolph Emil Kalman et al. A new approach to linear filtering and prediction problems. *Journal of basic Engineering*, 82(1):35–45, 1960. 29

Gregor Kastner. Sparse bayesian time-varying covariance estimation in many dimensions. *arXiv preprint arXiv:1608.08468*, 2016. 17, 47

Sangjoon Kim, Neil Shephard, and Siddhartha Chib. Stochastic volatility: likelihood inference and comparison with arch models. *The Review of Economic Studies*, 65(3):361–393, 1998. 46

Keith Knight and Wenjiang Fu. Asymptotics for lasso-type estimators. *Annals of statistics*, pages 1356–1378, 2000. 24

Chenlei Leng and Bo Li. Forward adaptive banding for estimating large covariance matrices. *Biometrika*, 98(4):821–830, 2011. 4, 59, 63, 64, 65

Elizaveta Levina, Adam Rothman, Ji Zhu, et al. Sparse estimation of large covariance matrices via a nested lasso penalty. *The Annals of Applied Statistics*, 2(1):245–263, 2008. 4, 59, 62, 63, 64, 65

Qing Li, Nan Lin, et al. The bayesian elastic net. *Bayesian Analysis*, 5(1):151–170, 2010. 14

Robert B Litterman et al. Techniques of forecasting using vector autoregressions. Technical report, 1979. 49

Hedibert F Lopes, Robert E McCulloch, and Ruey S Tsay. Parsimony inducing priors for large scale state-space models. 2014. 40, 43, 44

Alan Miller. *Subset selection in regression*. CRC Press, 2002. 18

Toby J Mitchell and John J Beauchamp. Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032, 1988. 2, 19

Trevor Park and George Casella. The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008. 4, 7, 13, 14

Giovanni Petris, Sonia Petrone, and Patrizia Campagnoli. Dynamic linear models. In *Dynamic Linear Models with R*, pages 31–84. Springer, 2009. 29, 38

Mohsen Pourahmadi. *High-Dimensional Covariance Estimation*. John Wiley & Sons, 2013. 60

Veronika Ročková. Sparse autoregressive processes for dynamic variable selection. 2016. 56, 57

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996. 1, 2, 4, 7, 11, 12, 13

Mike West. On scale mixtures of normal distributions. *Biometrika*, 74(3):646–648, 1987. 7, 13

Mike West and Jeff Harrison. *Bayesian Forecasting and Dynamic Models (2Nd Ed.)*. Springer-Verlag New York, Inc., New York, NY, USA, 1997. ISBN 0-387-94725-6. 29, 34, 38

Joe Whittaker. *Graphical models in applied multivariate statistics*. Wiley Publishing, 2009. 60

Max A Woodbury. Inverting modified matrices. *Memorandum report*, 42:106, 1950. 36, 37

Wei Biao Wu and Mohsen Pourahmadi. Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika*, 90(4):831–844, 2003. 61, 62

Yaming Yu and Xiao-Li Meng. To center or not to center: That is not the question—an ancillarity–sufficiency interweaving strategy (asis) for boosting mcmc efficiency. *Journal of Computational and Graphical Statistics*, 20(3):531–570, 2011. 48

Arnold Zellner. On assessing prior distributions and bayesian regression analysis with g-prior distributions. *Bayesian inference and decision techniques: Essays in Honor of Bruno De Finetti*, 6:233–243, 1986. 10

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006. 12

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005. 4, 7, 11, 12, 14